Current Paper: PastML

A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios

Sohta A. Ishikawa,[†,1,2,3] Anna Zhukova,[†,1] Wataru Iwasaki,[2] and Olivier Gascuel[*,1]
[1]Unité Bioinformatique Evolutive, Institut Pasteur, C3BI USR 3756 IP & CNRS, Paris, France
[2]Department of Biological Sciences, The University of Tokyo, Tokyo, Japan
[3]Evolutionary Genomics of RNA Viruses, Virology Department, Institut Pasteur, Paris, France

Diana Lin[1,2]

October 31, 2019

[1] Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC, Canada
[2] Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC, Canada

Outline

1. Background
    a. Review of Maximum Likelihood (ML)
    b. Introduction to Ancestral Scenario Reconstruction (ACR)
2. Methods
    a. Computing Marginal Posterior Probabilities (MPP)
3. Results
    a. Phylogeography of Dengue Virus
4. Applications
    a. Current uses of PastML
5. Conclusions
    a. General takeaways from PastML and its paper

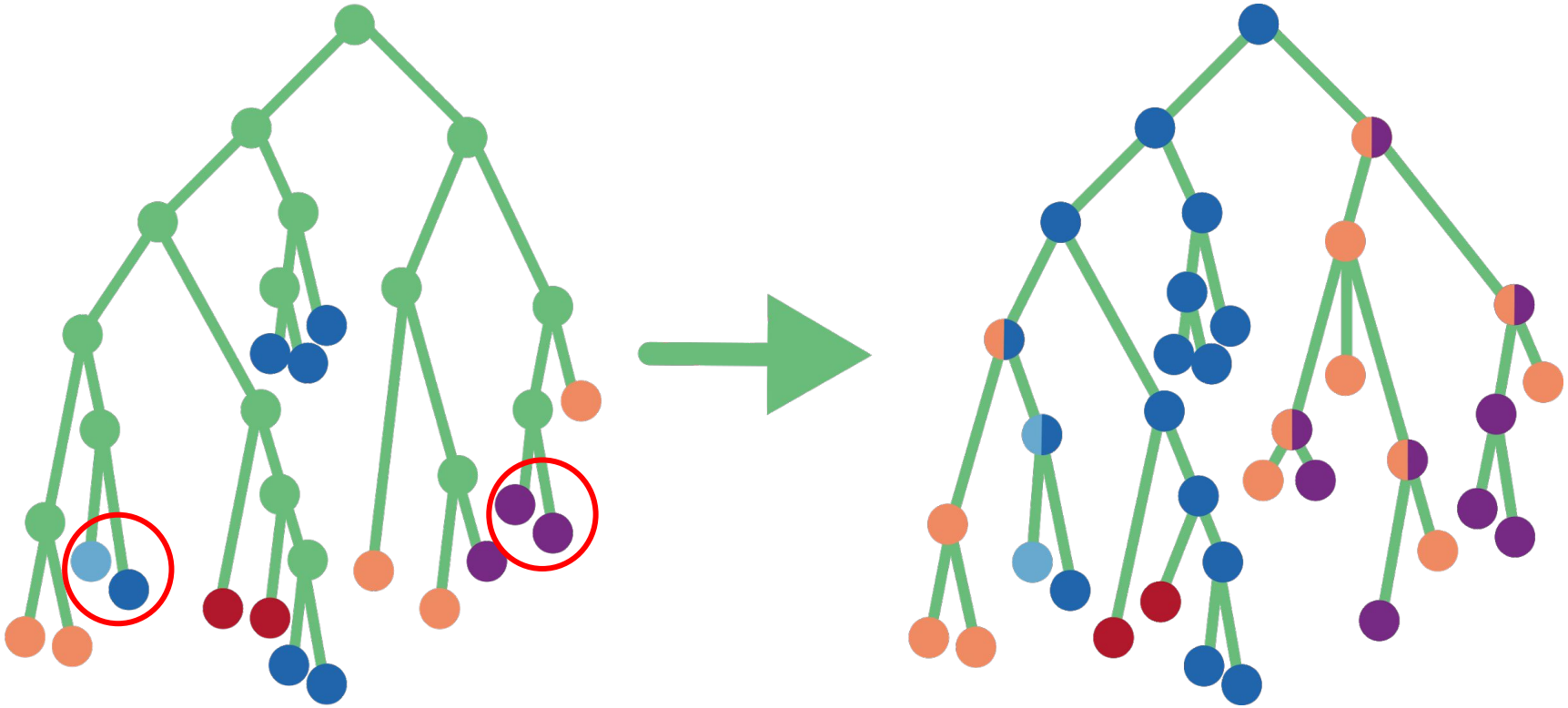
Background

Review From Last Week

- Maximum Likelihood
  - Choose the tree that maximizes the likelihood of the sequences to have evolved from the tree

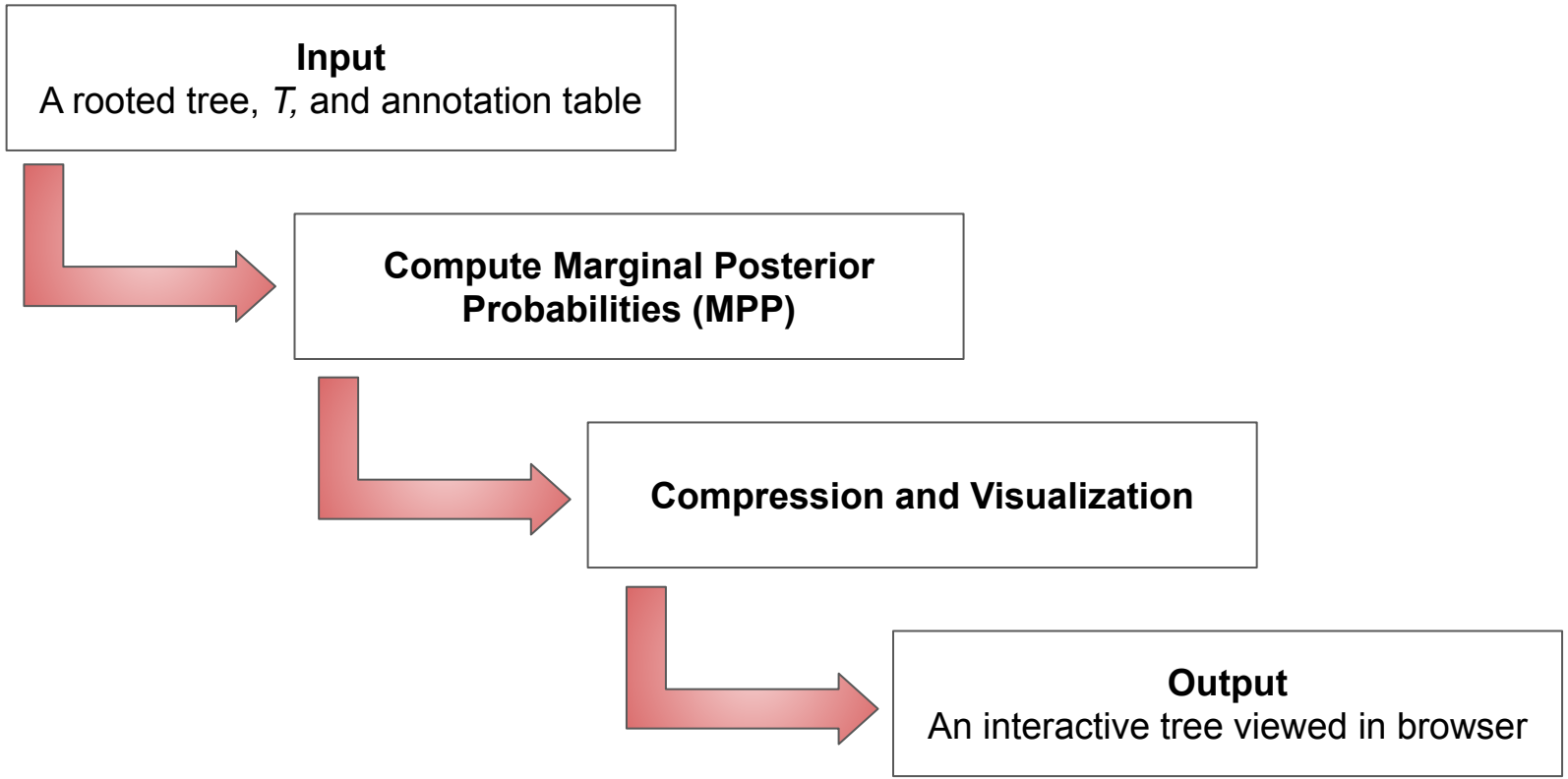
- PastML uses a variant of Maximum Likelihood to reconstruct ancestral scenarios
  - Choose the tree that maximizes the likelihood of the current character states to have evolved from the ancestral states of the tree

Ancestral Character Reconstruction (ACR)

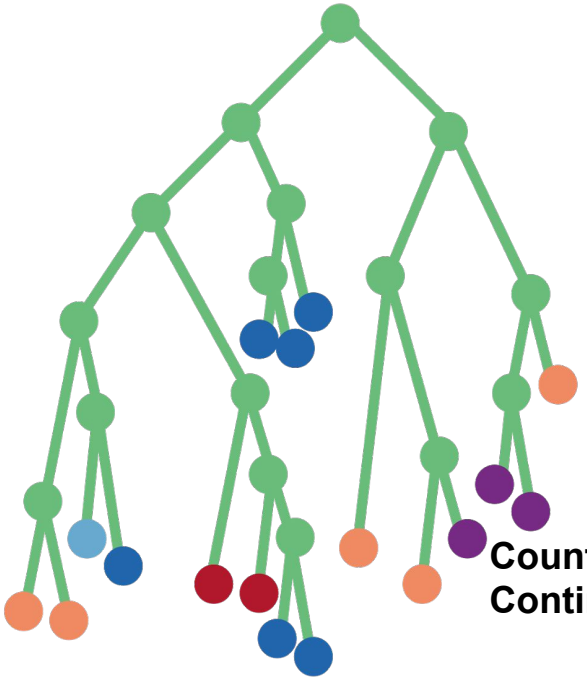


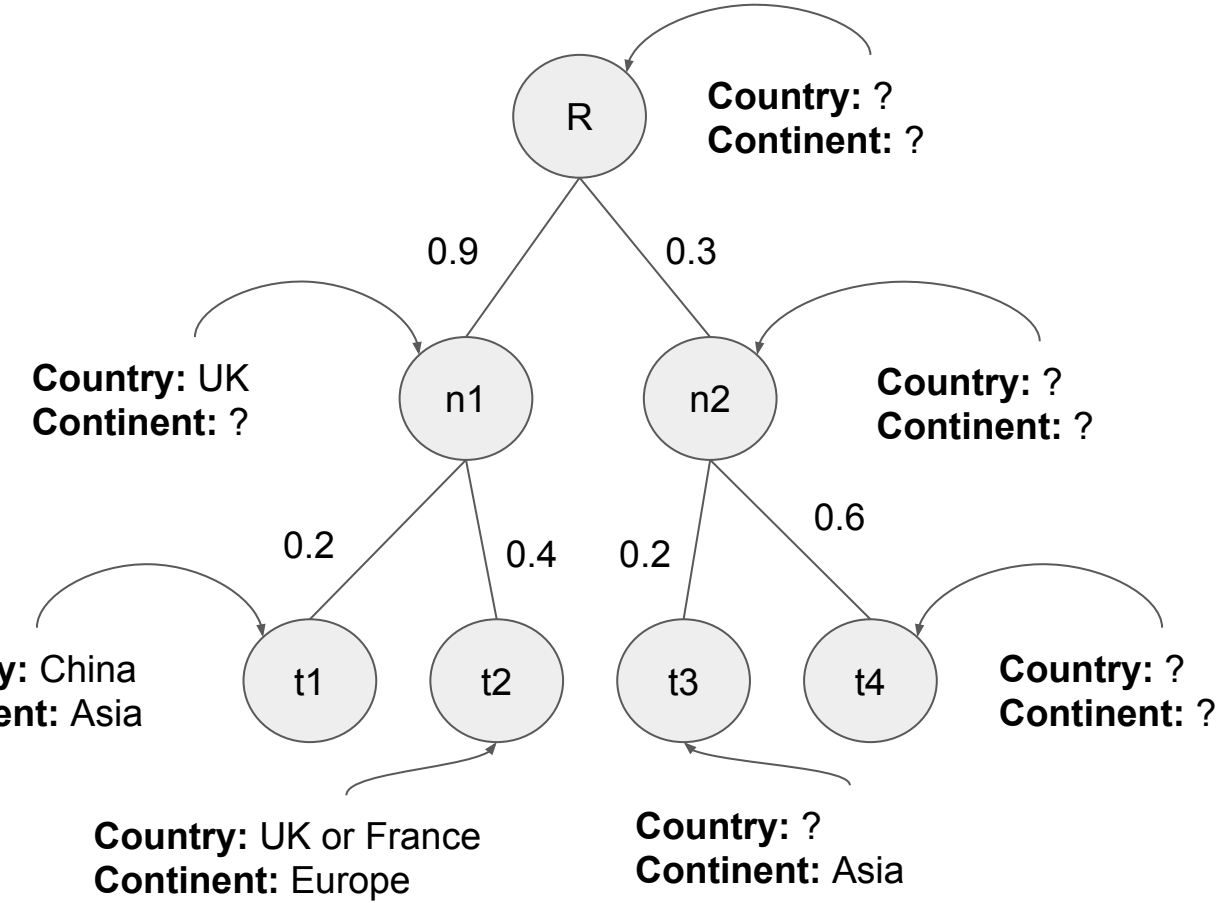
Methods

Overview

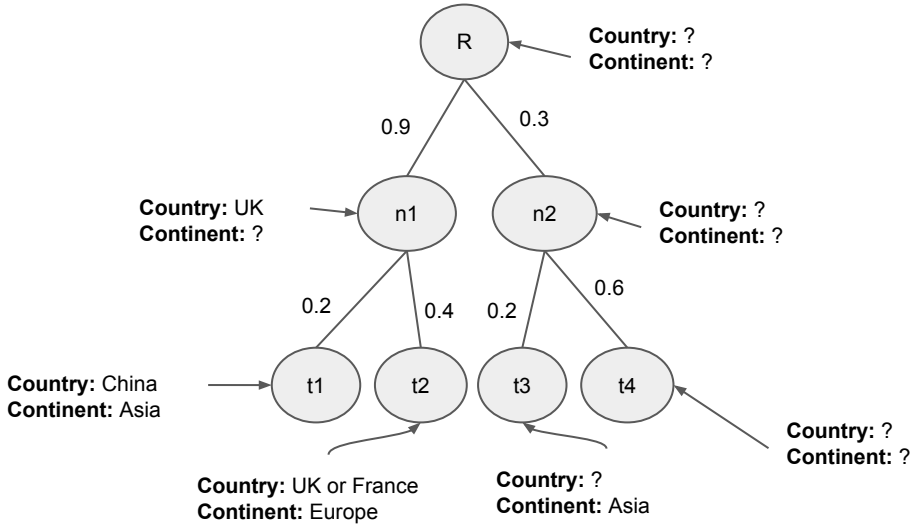

Input
A rooted tree, *T,* and annotation table
Compute Marginal Posterior Probabilities (MPP)
Compression and Visualization
Output
An interactive tree viewed in browser

Input




R
Country: ?
Continent: ?
0.9
0.3
Country: UK
Continent: ?
n1
n2
Country: ?
Continent: ?
0.2
0.4
0.2
0.6
Country: China
Continent: Asia
t1
t2
t3
t4
Country: ?
Continent: ?
Country: UK or France
Continent: Europe
Country: ?
Continent: Asia

Input



- Newick format tree:

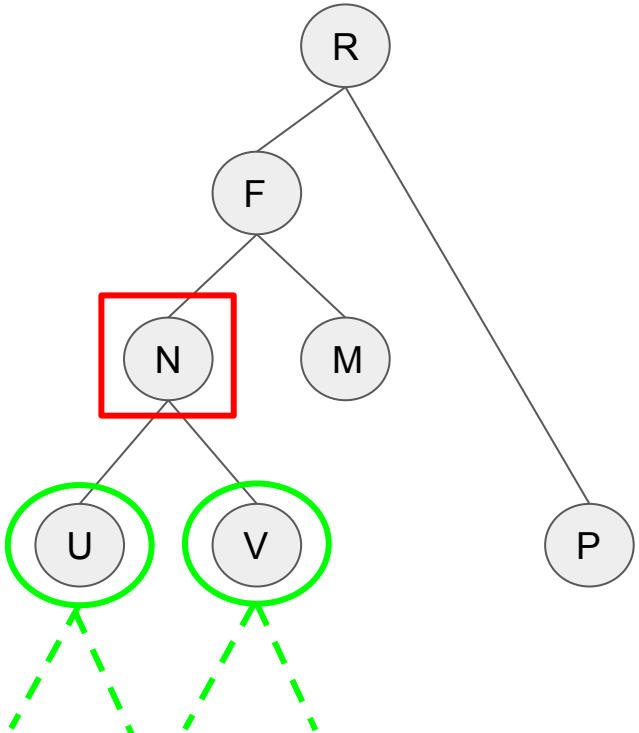
```
(t1:0.2,t2:0.4)n1:0.9,(t3:0.2,t4:0.6)n2:0.3)root
```

- Annotation table:

ID	Country	Continent
t1	China	Asia
t2	UK	Europe
t2	France	Europe
t3		Asia
n1	UK	

Computing Marginal Posterior Probabilities (MPP)

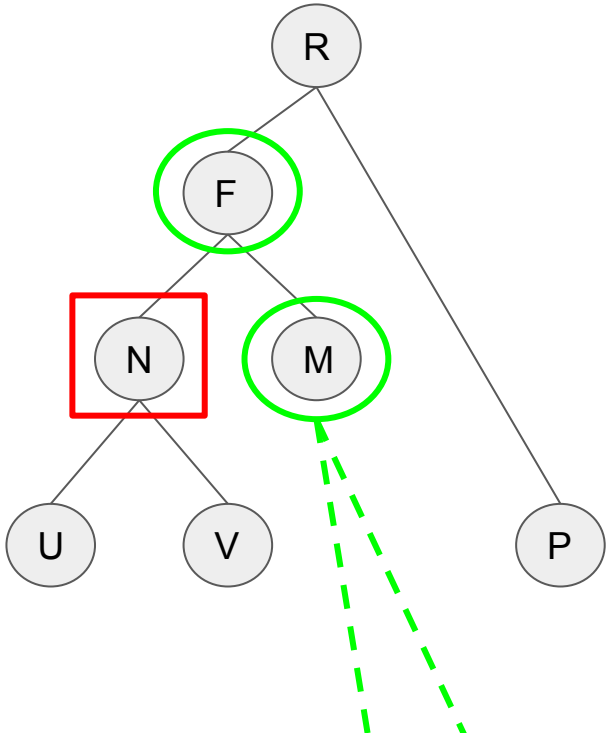
Given a tree, *T*, like this:



- Example: Node *N*
- **Down(*N*):** a vector of state likelihoods of Node *N* having state *i* given the “down” likelihoods
- Calculate all the Down() for each node of the tree (post-order traversal)
- **Recursive:** calculation of Down(N) is dependent on the calculation of Down() for all of N’s descendant nodes

Computing Marginal Posterior Probabilities (MPP)

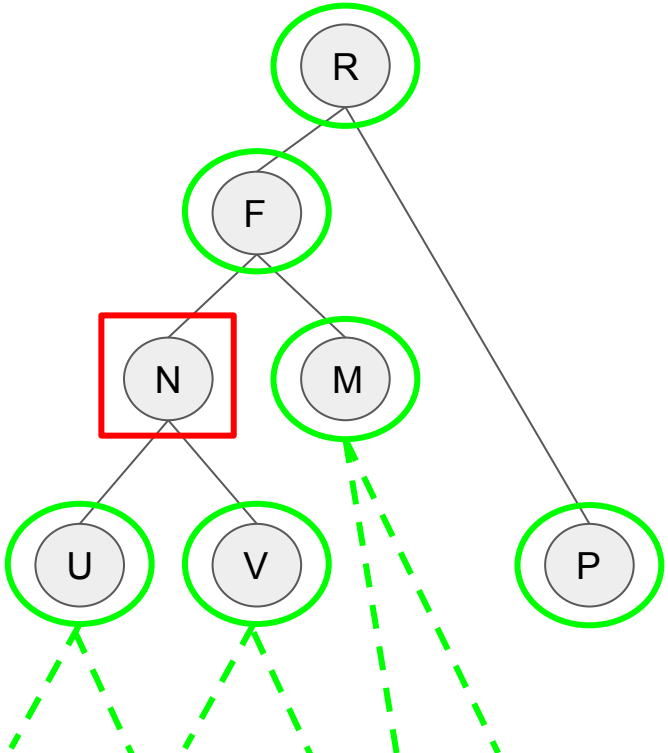
Given a tree, *T*, like this:



- Example: Node *N*
- **Up(*N,i*):** a vector of state likelihoods that Node *N* has state *i* given the “up” likelihoods
- Calculate all the Up() for each node of the tree (pre-order traversal)
- **Recursive:** calculation of Up(*N*) is dependent on the calculation of Up() for N’s parent node and Down() of its siblings
- That is, Up(*N*) is dependent on Up(*F*) and Down(*M*)

Computing Marginal Posterior Probabilities (MPP)

Given a tree, *T*, like this:



- Example: Node *N*
- **Marginal*(N,i)*:** the marginal posterior probabilities of Node *N* having state *i*

$$\text{Marginal}(N,i) = \frac{\pi_i \text{Down}(N,i)\text{Up}(N,i)}{\text{TotalProba}(N)}, \text{ where}$$

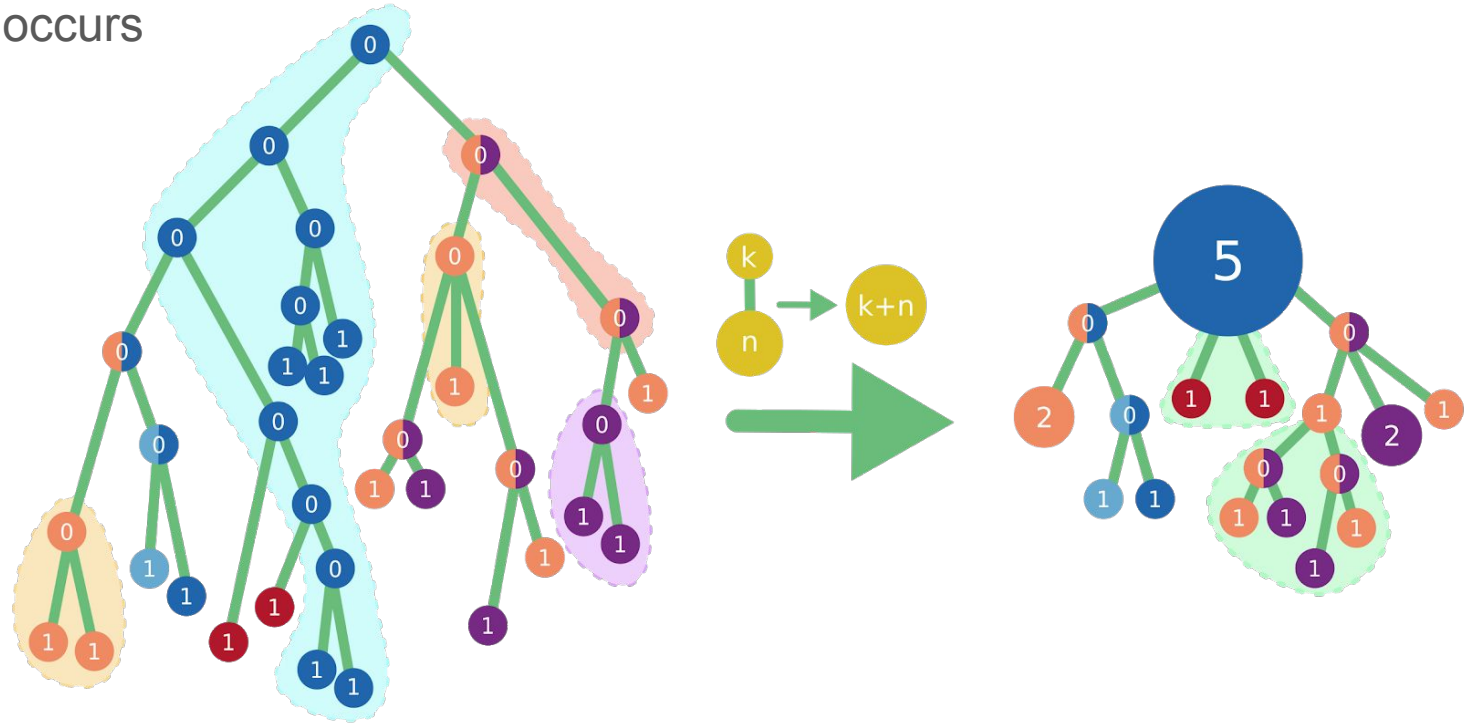
(law of total probability):

$$\text{TotalProba}(N) = \sum_{j \in S} \pi_j \text{Down}(N,j)\text{Up}(N,j).$$

- MPPs with a higher Brier score are discarded

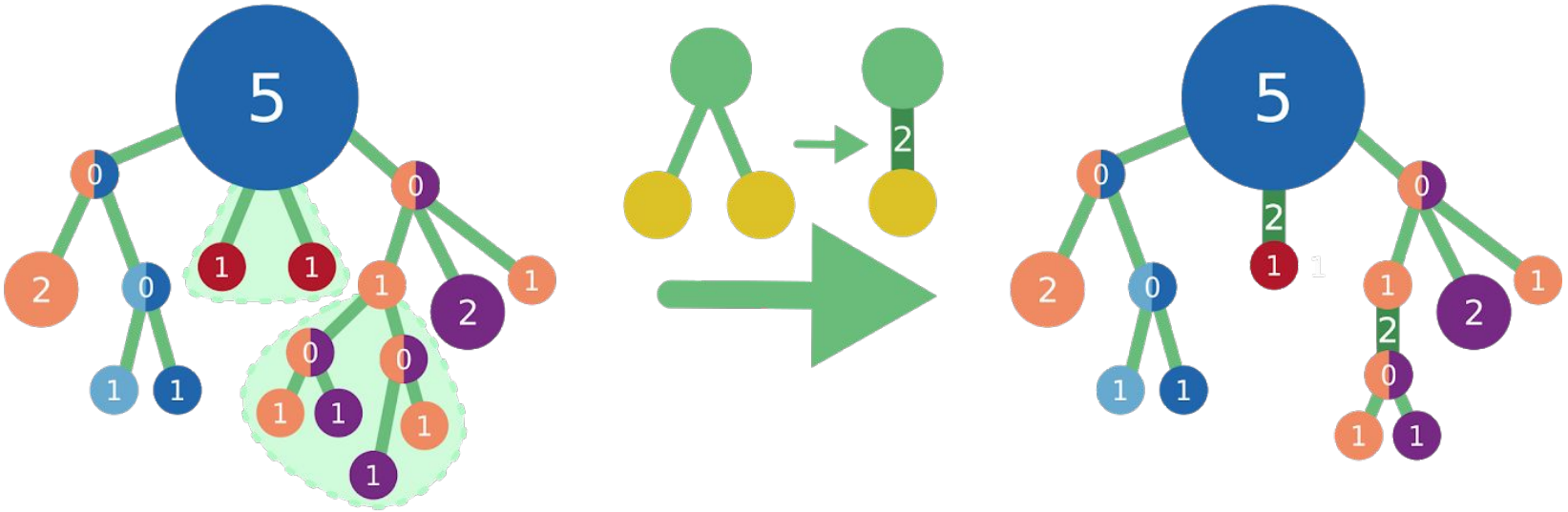
Compression and Visualization

- **Vertical Merge:** cluster together parts of the tree where no state change occurs



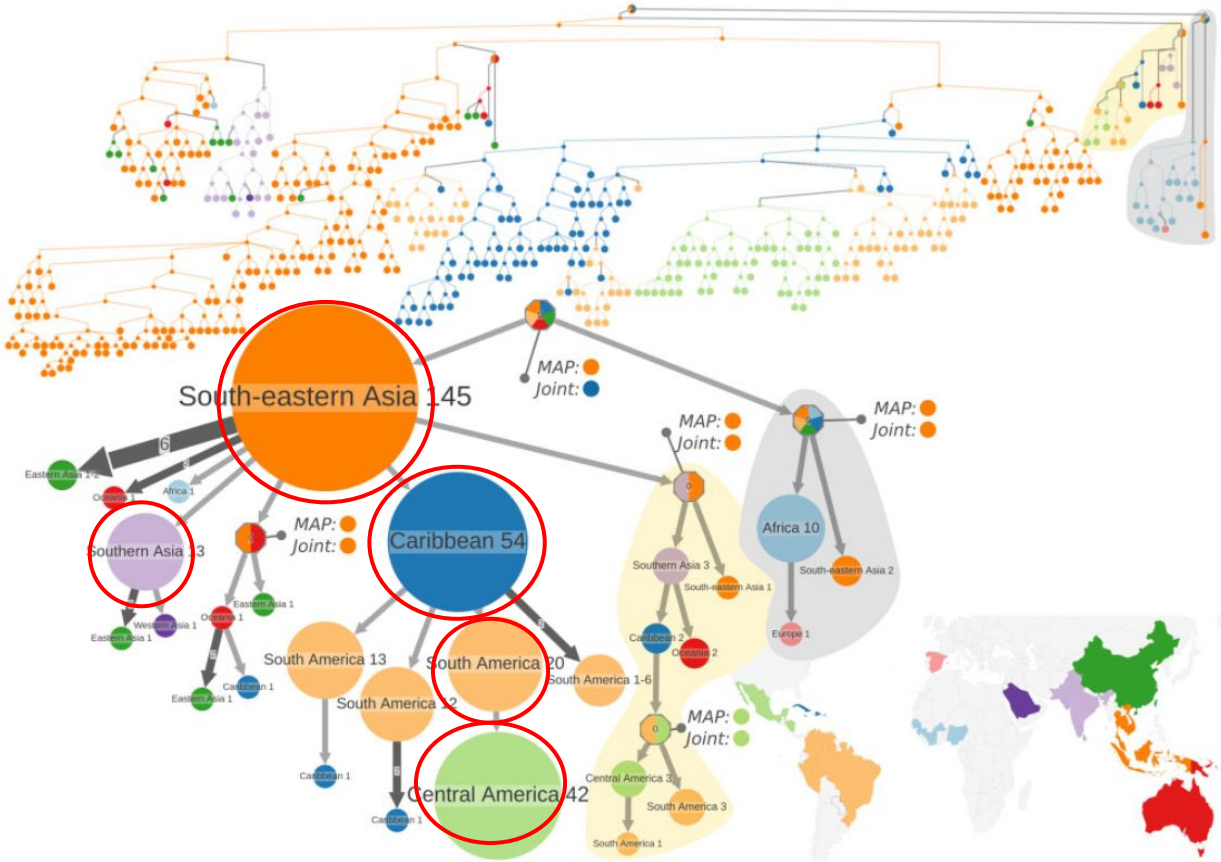
Compression and Visualization

- **Horizontal Merge:** clusters independent events of the same kind



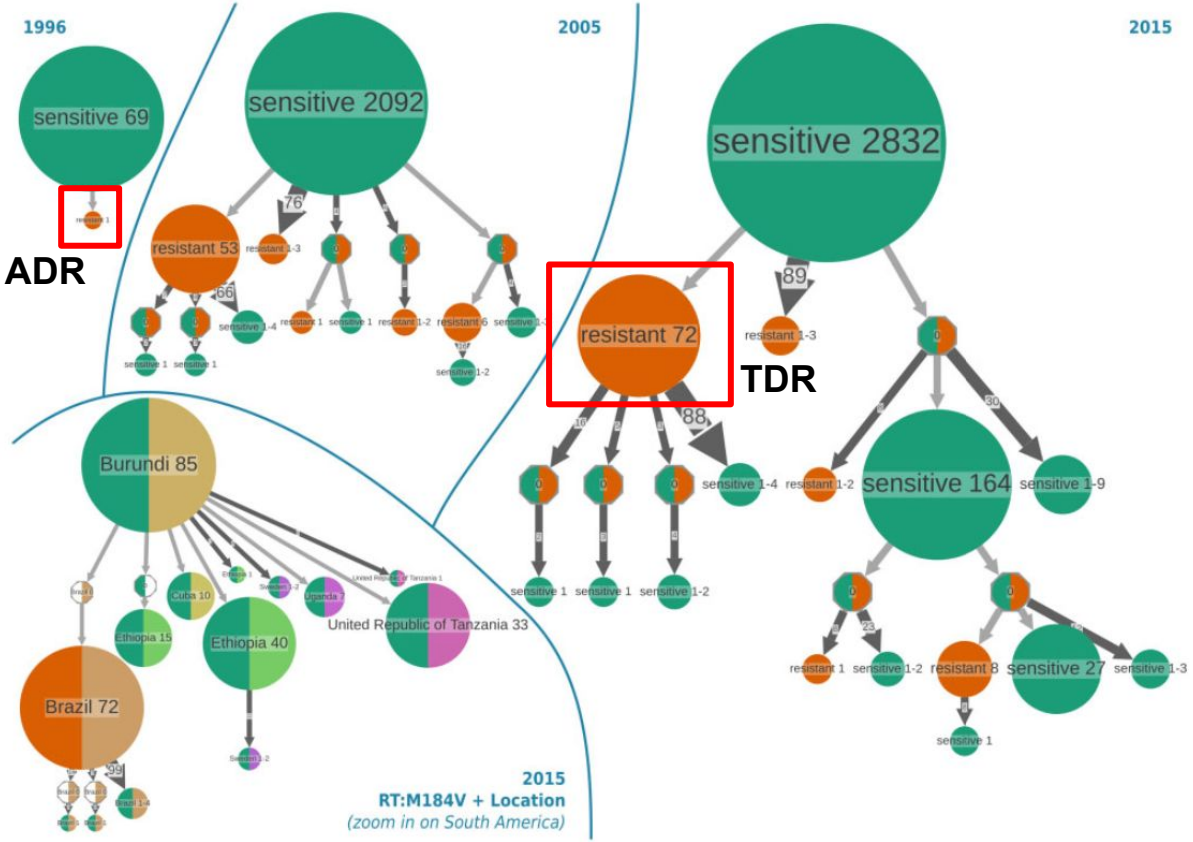
Results

Phylogeography of Dengue Epidemics



- Constructed from medium-sized dataset of 356 sequences
- Resolved main transmission route and shows uncertainty of root

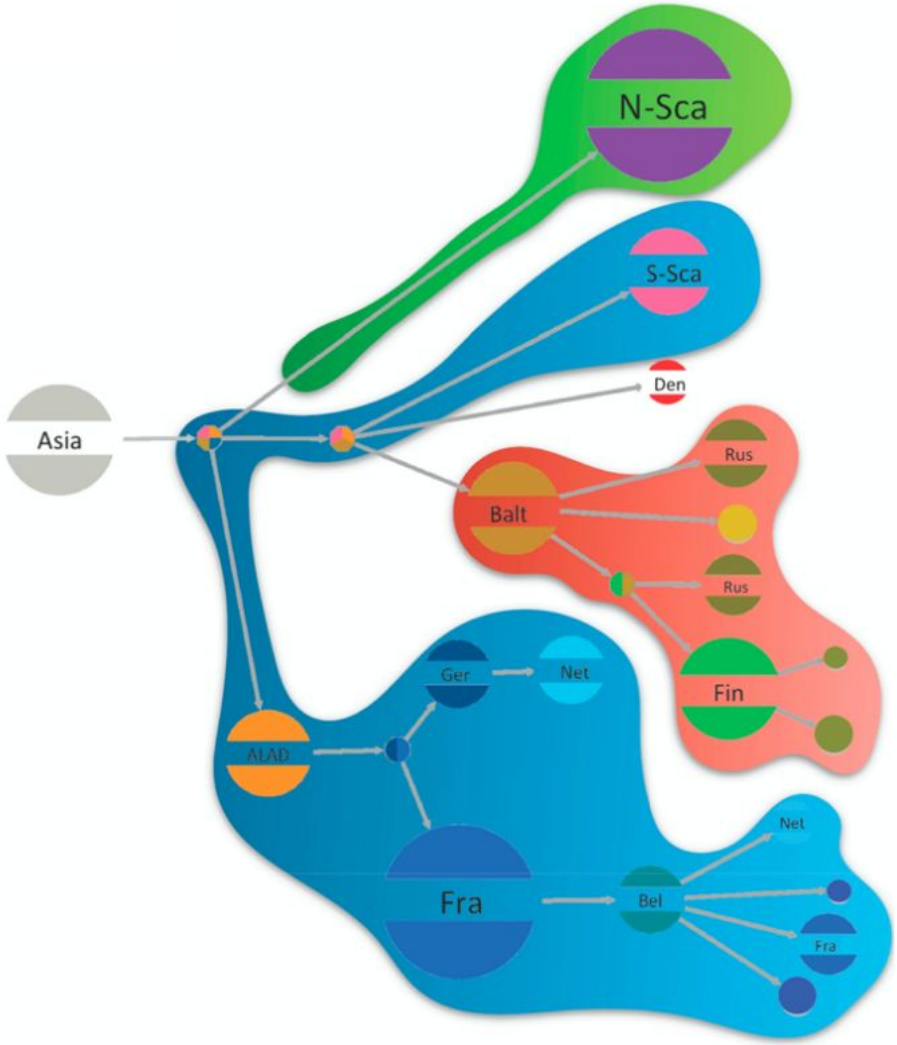
Drug Resistant Mutations in HIV



- Constructed from a dataset of 3,619 HIV-C *pol* sequences
- **Acquired Drug Resistance (ADR)** represented by single resistant node
- **Transmitted Drug Resistance (TDR)** represented by cluster of resistant nodes

Applications

- Castel et. al. **Phylogeography of Puumala orthohantavirus in Europe.** *Viruses.* 24 July 2019
- Reconstruct the phylogeographical spread of modern PUUV throughout Europe during the last postglacial period
- Three potential dispersal routes of PUUV identified



Pros

- New, simple, fast
- Many user-specified parameters for the tool
- Online interface https://pastml.pasteur.fr is easy for everyone to use

Cons

- Requires a pre-constructed tree using other phylogeny construction software
  - Most phylogeny tools generate an **unrooted** tree that cannot be used as input to PastML without being rooted

Conclusions

- ➢ PastML is used for...
  - ✓ Reconstructing **AND** visualizing ancestral scenarios
- ➢ PastML is accurate and robust because…
  - ✓ It can be used on large data sets (while maintaining its speed)
  - ✓ Its results are in consensus with previous studies
- ➢ PastML is impactful for the field of phylogenetics because…
  - ✓ Reconstructing ancestral scenarios can lead to more accurate predictions (especially important in the study of viral mutations)

Questions?

Appendix

Probability

- **Marginal:** the probability of event A occurring (unconditional - i.e. it is not conditioned on another event
  - Example: the probability of a red card drawn: 26/52 = 0.5
  - Example: the probability of a 4 card drawn: 4/52 = 1/13
- **Joint**: the probability of event A and event B occurring (i.e. the probability of the intersection of two or more events)
  - Example: the probability that a red 4 card is drawn: 2/52 = 1/26
- **Conditional**: the probability of event A occurring, given that event B occurs
  - Example: given that you drew a red card, what's the probability that it's a 4: 2/26 = 1/13

https://sites.nicholas.duke.edu/statsreview/jmc/

Prior vs Posterior

- **Prior**: probability distribution before you have sampled an data and attempted to estimate the character of interest
  - Denoted as $\pi(\theta)$
- **Posterior**: probability distribution after you have sampled the data (conditional, since it depends on observed data)
  - Denoted as $\pi(\theta|X)$
- **Bayes' Theorem:** $$\pi(\theta|X) = \frac{f(X|\theta)\pi(\theta)}{\sum_{\theta_i} f(X|\theta_i)\pi(\theta_i)}$$

http://www.stat.cmu.edu/~brian/463-663/week09/Chapter%2003.pdf

Methods Summary

- Maximum Parsimony
  - Choose the tree that has the fewest character changes
- Maximum Likelihood
  - Choose the tree that maximizes the likelihood that the sequences evolved from the given tree
- Bayesian Inference
  - Like ML, but instead of choosing for a single tree, a sample is taken of a large number of trees with high likelihoods
- Markov Chain Monte Carlo
  - Start with a trial tree, and changes it slightly, if this change improves the likelihood, this is is the next tree in the sample - consensus is made of all the sample trees

http://www.gen-info.osaka-u.ac.jp/~yonishi/bioinfo-new.pdf

F81-like Model

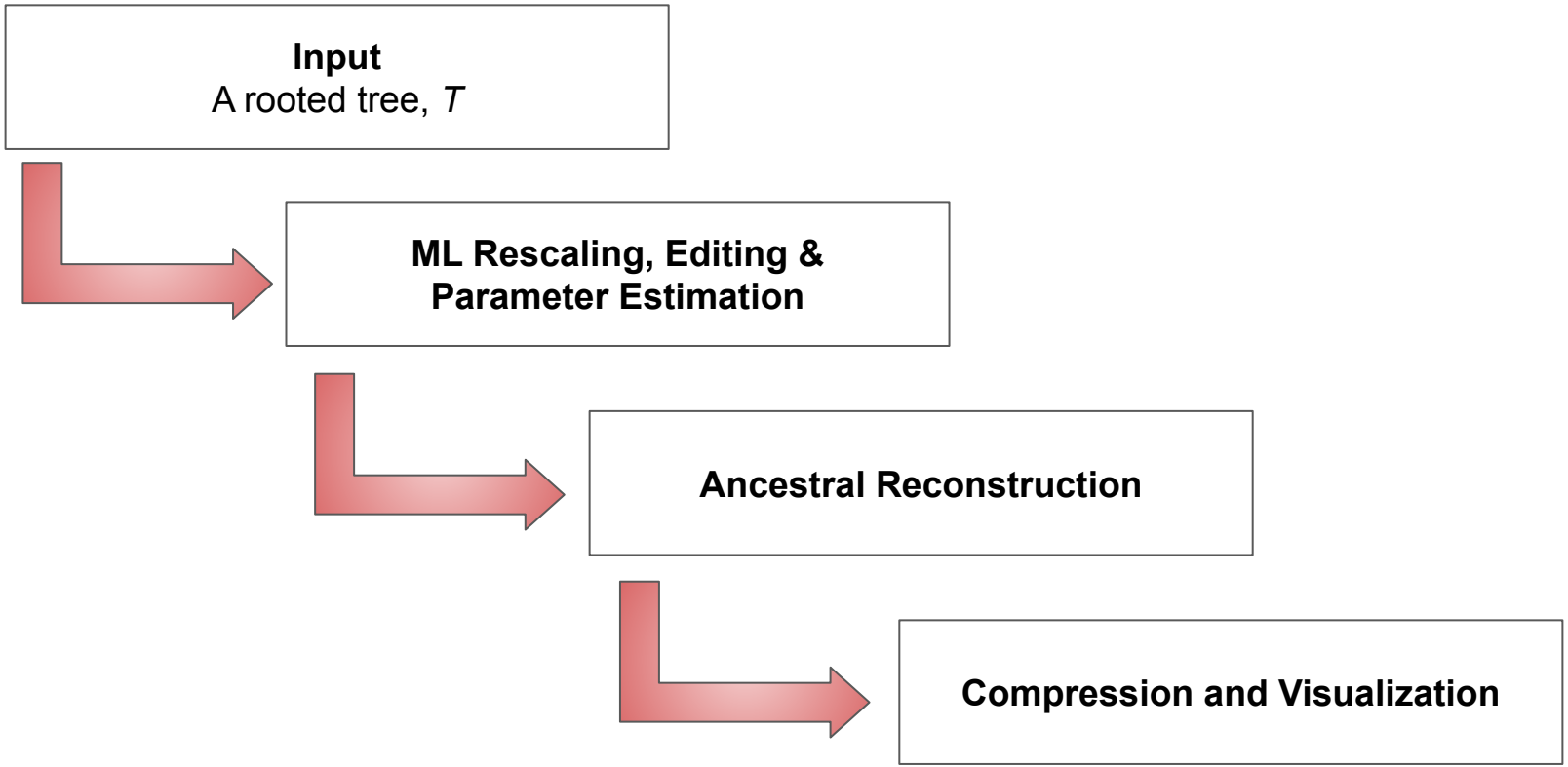
- the rate of changes from i to j (i ≠ j) is proportional to the equilibrium frequency of j
- the probability of changes along a branch of length t is simply expressed as:

$$\begin{aligned} PC(i \to j/t) &= (1 - e^{-\mu t})\pi_j \qquad \text{if } j \neq i \\ &= e^{-\mu t} + (1 - e^{-\mu t})\pi_i \quad \text{otherwise,} \end{aligned}$$

where μ is the normalization factor:

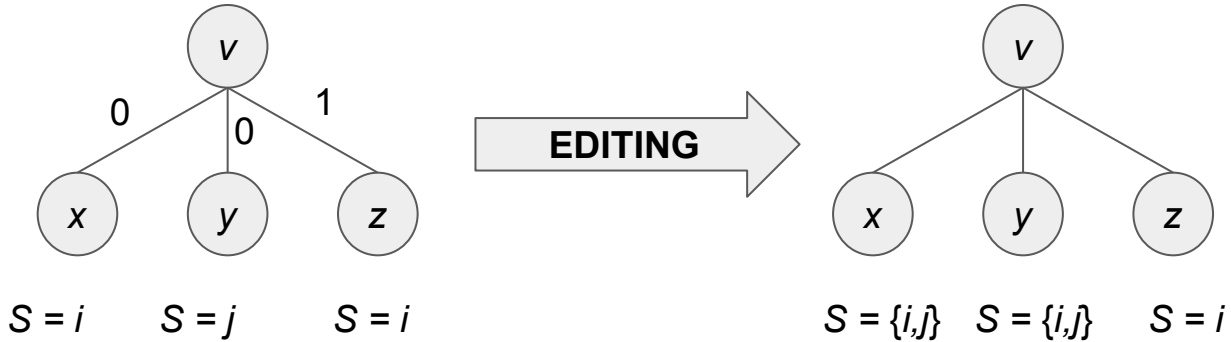
$$\mu = 1/(1 - \sum \pi_i^2).$$

Overview


Input
A rooted tree, *T*
ML Rescaling, Editing & Parameter Estimation
Ancestral Reconstruction
Compression and Visualization

ML Rescaling, Editing, and Parameter Estimation

- Rescaling: The number of character changes along the tree is proportional to the branch lengths of the input tree
- Editing: Some internal branches have a length of zero, so the union of these states are computed and assigned to these tips



- Parameter estimation using ML

Computation of the Marginal Posterior Probabilities

$$\text{Down}(N,i) = [\sum_j \text{PC}(i \to j/u)\text{Down}(U,j)]$$

$$\times [\sum_j \text{PC}(i \to j/v)\text{Down}(V,j)],$$

and for a tip l : if $c(l) = i$ or X, then $\text{Down}(l,i) = 1$,

else $\text{Down}(l,i) = 0$.

Computation of the Marginal Posterior Probabilities

$$\mathrm{Up}(U,i) = \left\{ \sum_j \mathrm{PC}(i \to j/u)\mathrm{Up}(N,j) \right.$$
$$\left. \times \left[\sum_k \mathrm{PC}(j \to k/v)\mathrm{Down}(V,k) \right] \right\}$$

Brier Criterion

- A score to measure the accuracy of predicted probabilities

$$\mathrm{Brier}(N) = \sum_{i \in S} \underbrace{[\mathrm{PPr}(N,i)}_{0 \le \mathrm{PPr}(N,i) \le 1} - \underbrace{\mathrm{Truth}(N,i)}_{0 \text{ or } 1}]^2$$

- The *lower* the Brier score, the *better* the prediction

Marginal Posterior Probabilities Approximation

- This method returns the “best” set of likely states per node
- If the length of this set is *k*, then each state in this set has a probability of 1/*k*
  - Discarded states will have a probability of 0
- The set is then reordered from highest marginal posterior probability to lowest
- Select the states that have the smallest Euclidean distance between Marginal(*N*) and the probability vector defined by {1/*k*, … }
- This vector is known as the MPPA

Ultimately, the algorithm minimizes the difference between the square root of the Brier scores of MPPA and MPP for each node.

$$\sqrt{\text{BrierMPPA}(N)} - \sqrt{\text{BrierMPP}(N)} \leq D_k(N)$$

ML-based Methods

- **Joint:** infers the most likely ancestral scenario over all the tree and possible state values
  - Returns one unique state for each node (joint estimation of the most likely state)
- **Marginal:** computes the marginal likelihoods and posterior probabilities of all states for all internal nodes
  - Returns all states assigned with a probability
- **Maximum a posteriori (MAP) :** computes the marginal posterior probabilities for every state for each tree node, based on information from the whole tree (tip states, branch lengths, etc)
  - Returns the state with the highest posterior for each node, independent of other nodes
- **MPPA:** approximates the state posteriors
  - Return for a subset of likely states for every node

https://pastml.pasteur.fr/help

Future Directions

- More maximum likelihood algorithms to choose from
- More evolutionary models to choose from
- More scoring rules to choose from
- Refinement to provide users with a global view of evolutionary processes
  - Strain flow between regions of countries
  - Acquisitions and losses of molecular characters
  - Dynamics of ecological character changes
- Compare multiple ancestral scenarios and produce a consensus