# Redelings & Suchard, “Joint Estimation of Alignment and Phylogeny”

Michael Nute

4/5/2016

# Background: Alignment/Tree Estimation

Challenge: Phylogeny estimation depends on an alignment, but alignments are improved with the use of a guide tree.

- Possible solutions:
  - Construct the alignment without the guide tree.
    - Paper suggests potentially removing ambiguous regions of the alignment, but this is not recommended today.
  - Construct the tree with a basic alignment, then use that guide tree for a new alignment.
  - Iterate between estimating alignment and tree.
  - Co-estimate the alignment and tree based on a joint likelihood model.
    - ***This is what BAli-Phy Does***

# Background: Bayes Theorem

- Basic Bayesian Model:
  - Let $\theta$ be some parameter of interest. Assume that it has a *prior* distribution $\pi(\theta)$.
    - Prior can be based on our beliefs, or some separate study, or something else, but it's the estimate of the probabilities for $\theta$ *before* we see any data.
  - Let $X$ be some data of interest, and that $X$ comes from a distribution that is defined by $\theta$, i.e. $X \sim f(x|\theta)$
  - The *Bayesian Posterior* for $\theta$, given observation $x$ is:

$$\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{\int_{\Theta}\pi(\theta)f(x|\theta)d\theta}$$

  - Don't worry about the math though, the point is that once we observe some data, our beliefs about the original parameter can shift to give us a new set of probabilities for the parameter.
  - For many applications, this is an effective way let initial beliefs blend with observed data. Sometimes it works very well.

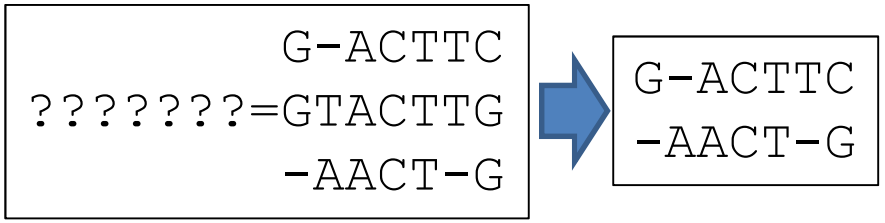
# Background: MCMC Sampling

MCMC=Markov-Chain Monte-Carlo

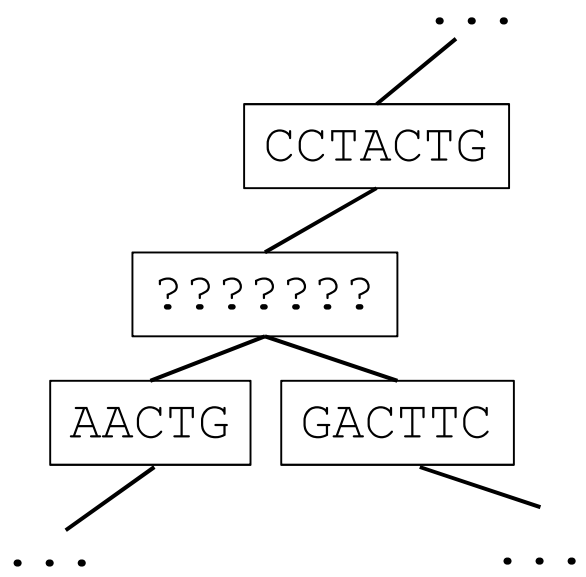
1. Monte Carlo Simulation:
   - Imagine you have a bunch of things that are random. Call them $X_1, X_2, \ldots, X_k$ and assume that they have pretty vanilla probability distributions: $X_1 \sim f_1$,…
   - Now imagine you're interested in some weird function of all those things, like $Y = (X_1 + X_2 + \sin(X_3\, X_2/X_4))/X_5$ or something. Suddenly the distribution of $Y$ is not so vanilla.
   - The easiest approach is to simulate all the $X$'s, calculate $Y$, then repeat many times. The distribution of results will get you a picture of the distribution of $Y$. Not exactly, but close enough.
2. Markov Chains:
   - Sometimes the only good way to do this is to simulate one component (e.g. $X_1$) at a time, leaving the others fixed, and tally the $Y$'s each time.
   - Eventually, the distribution of $Y$'s gets to the right place.

# Background: Two More Things

1. If we have a set of sequences and a phylogeny, if we know the alignment along each internal edge, then we know the whole alignment.

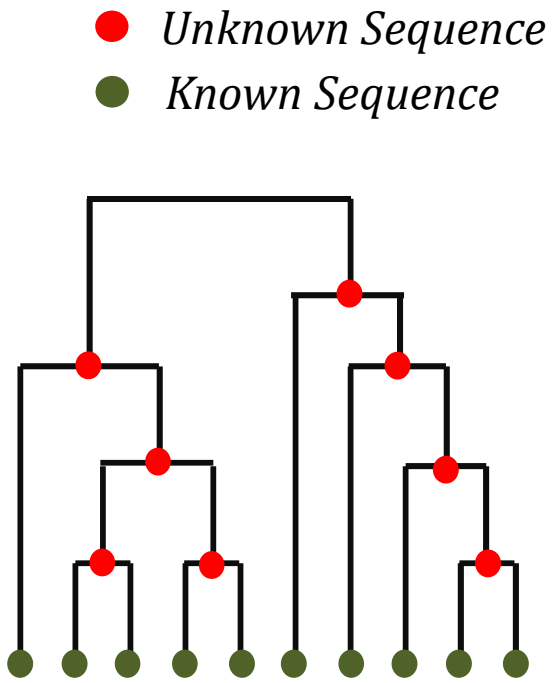



2. For an internal node, the likelihood of any given sequence at that node only depends on the values at the *three* sequences around it.
   - This is an implicit assumption we make about the way evolution works.



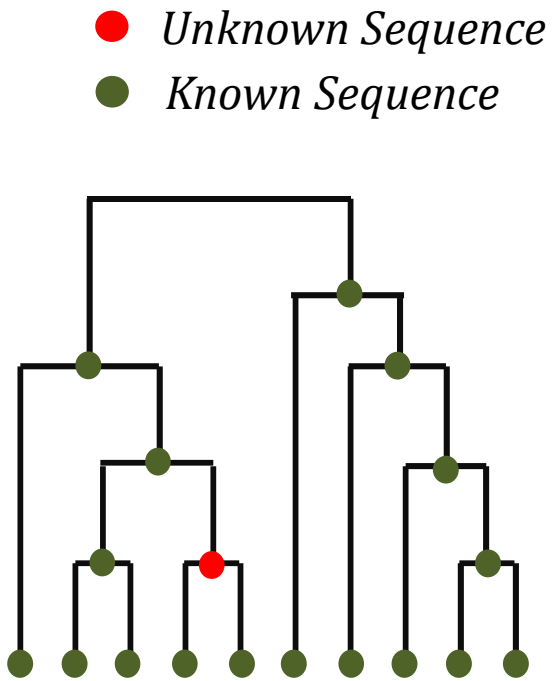
# BAli-Phy: How it Works

- Starts by specifying a Probability Model for sequence evolution, including with insertions and deletions, then specifying a ***prior*** for every parameter in the model, *including* the alignment and the tree topology.
- Crucial part: it chooses the priors and distributions so that it is easy to pick a new sequence for an internal node *with probability proportional to the likelihood given the three nodes around it.*
- That's the main step that the algorithm repeats over and over again to "sample" the alignment



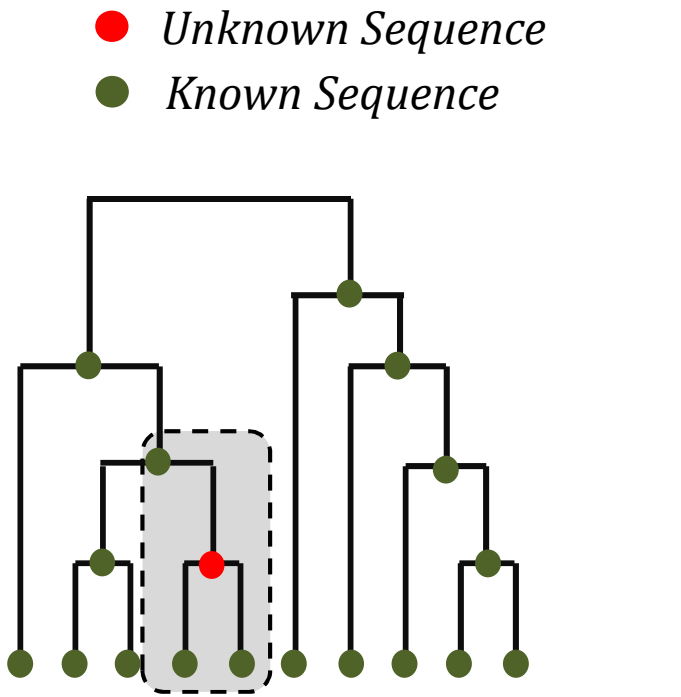
# BAli-Phy: Algorithm

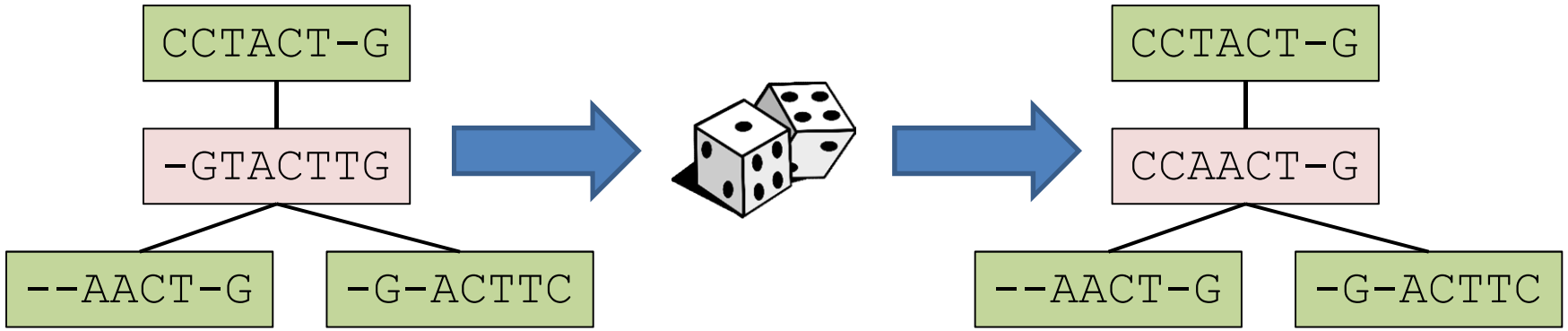
- Step 1:
  - Initialize every internal node to a sequence that is aligned with the three nodes around it.
  - Pick initial parameters from the prior distributions.
  - Pick a sequence to start with (red).



# BAli-Phy: Algorithm

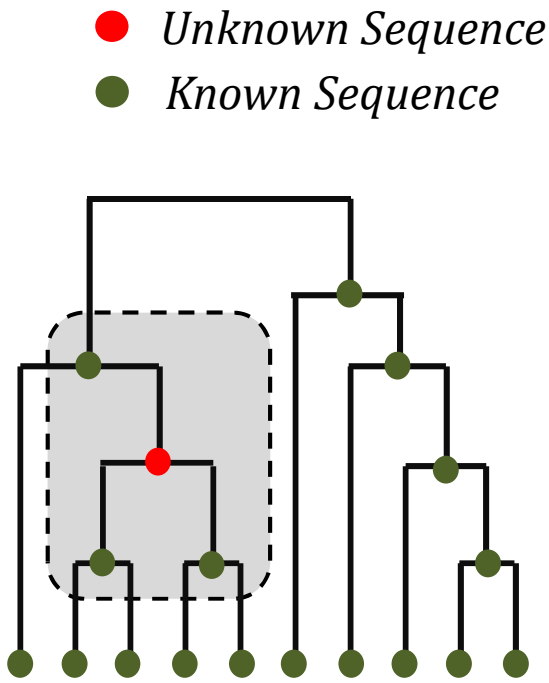
- Step 2a:
  - Choose a new sequence & alignment for the red node proportional to the conditional likelihood given the nodes around it.

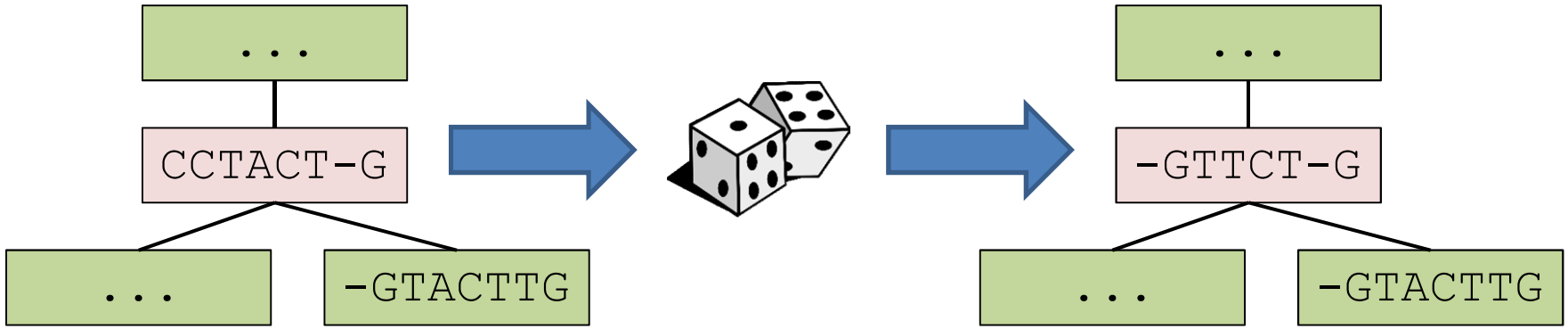


# BAli-Phy: Algorithm

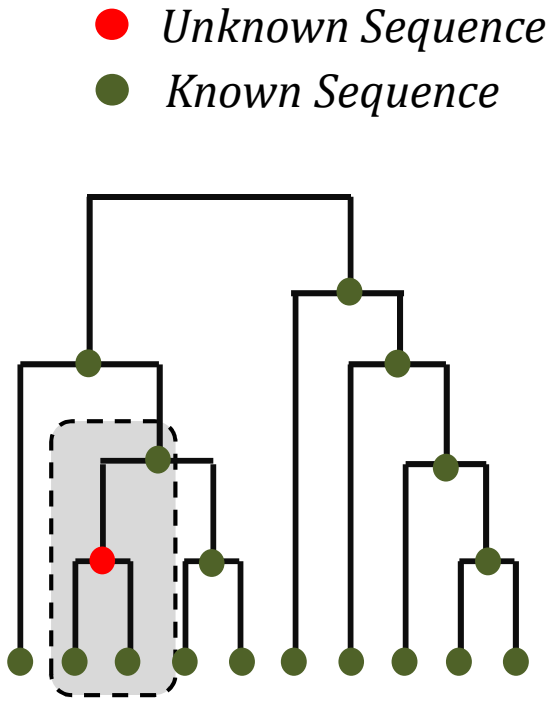
- Step 2b:
  - Step to the next sequence. (There may be more than one option.)
  - Do the same thing as last time

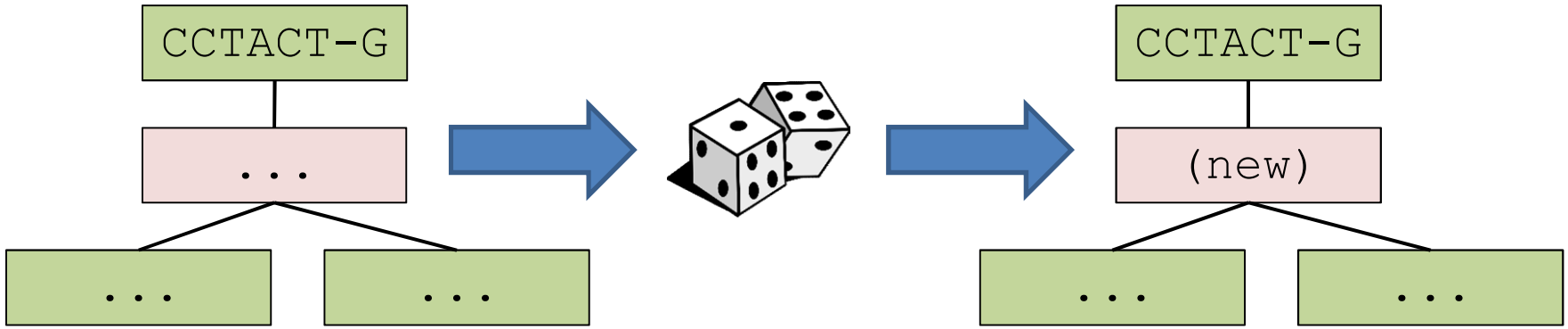


# BAli-Phy: Algorithm

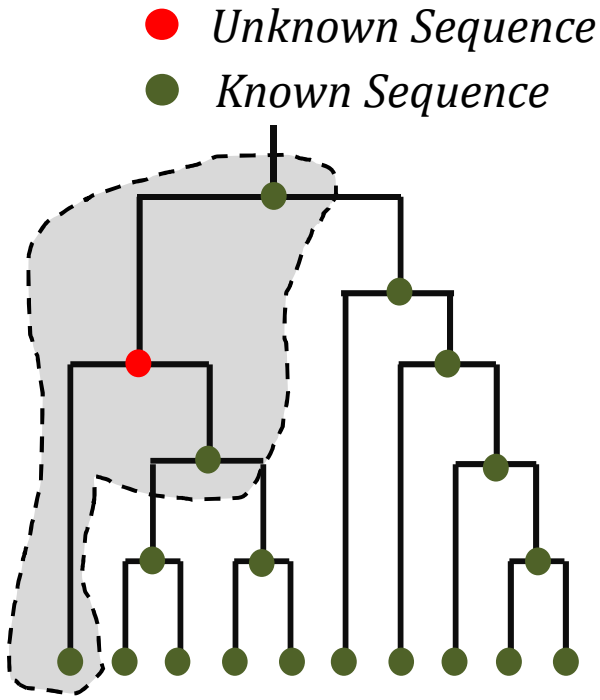
- Step 2(…):
  - Step to the next sequence. (There may be more than one option.)
  - Do the same thing as last time

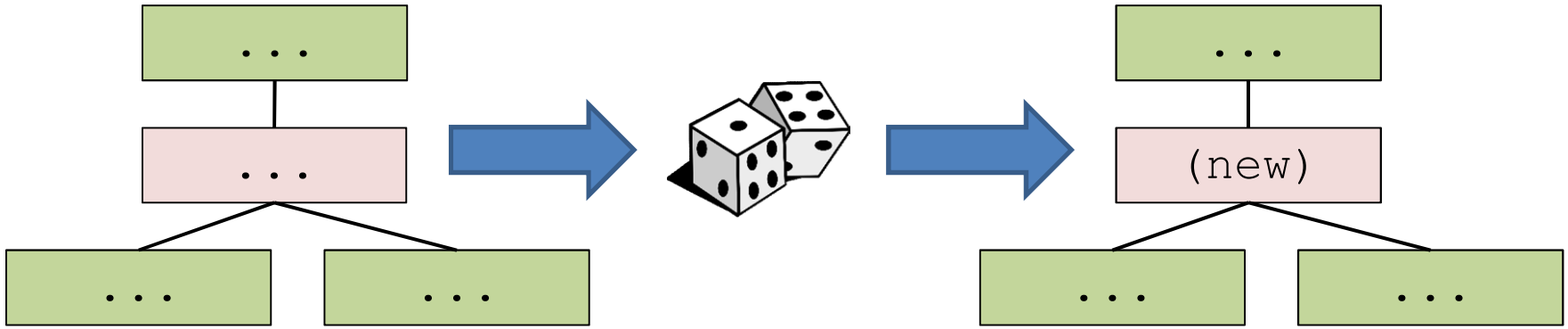


# BAli-Phy: Algorithm

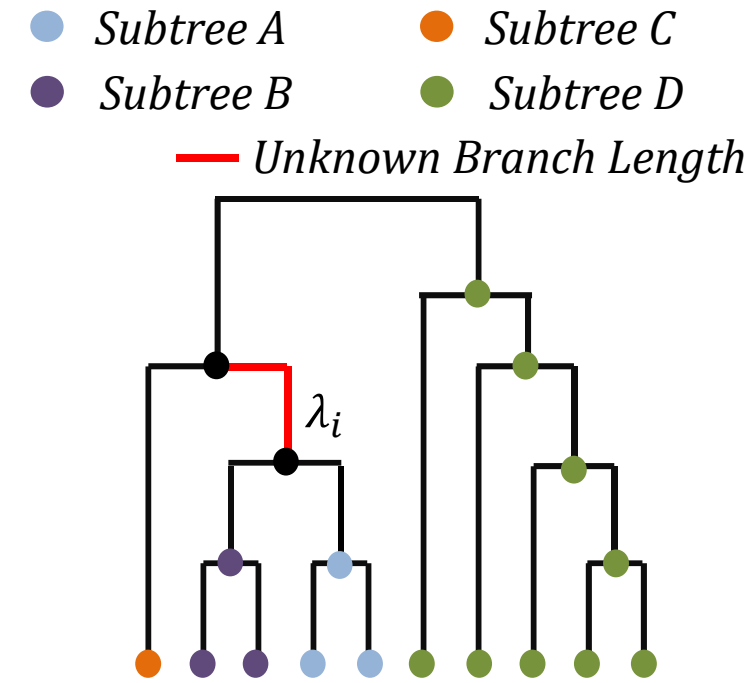
- Step 2(…):
  - Step to the next sequence. (There may be more than one option.)
  - Do the same thing as last time



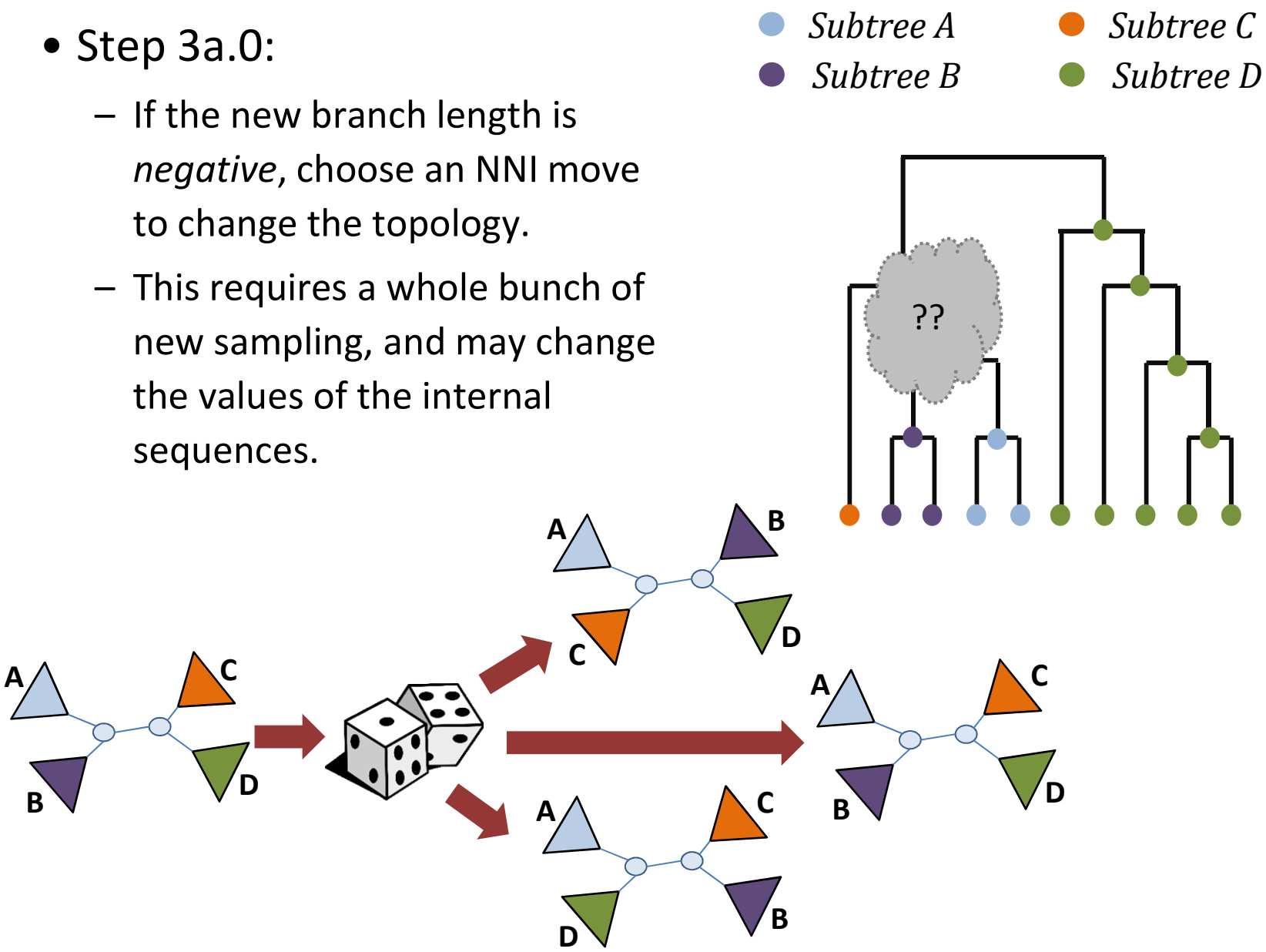
# BAli-Phy: Algorithm

- Step 3a:
  - Choose an internal branch to start with.
  - Re-sample its branch length, conditional on the sequences on either side of it.



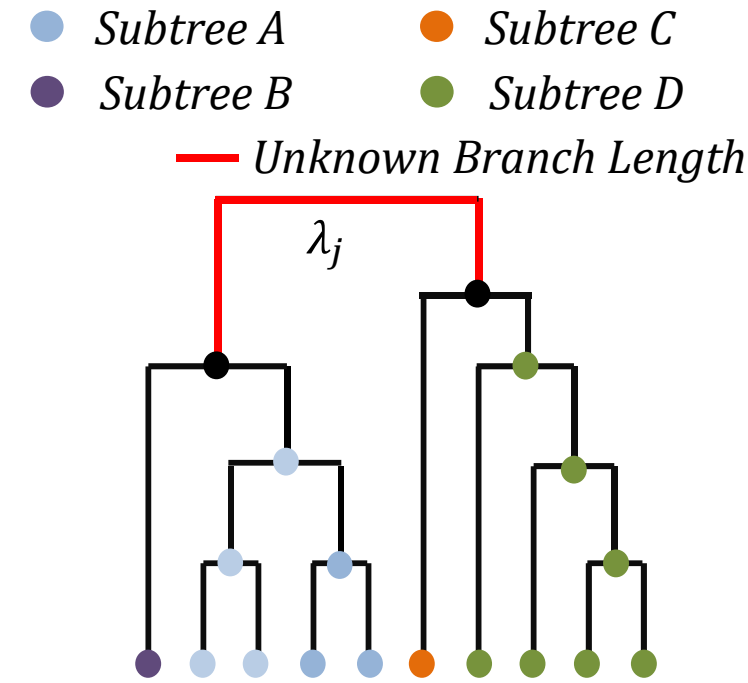

# BAli-Phy: Algorithm

- Step 3a.0:
  - If the new branch length is *negative*, choose an NNI move to change the topology.
  - This requires a whole bunch of new sampling, and may change the values of the internal sequences.



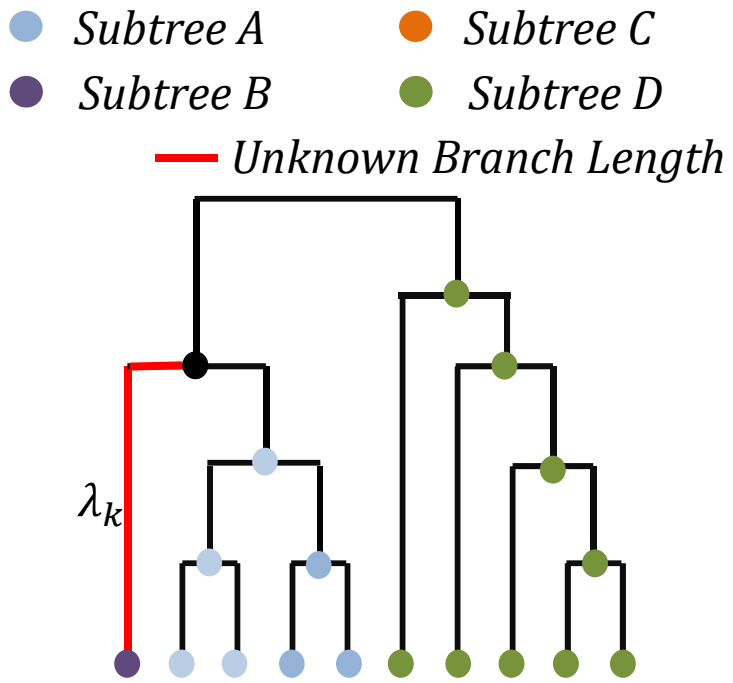
# BAli-Phy: Algorithm

- Step 3b:
  - Choose an internal branch to start with.
  - Re-sample its branch length, conditional on the sequences on either side of it.
- Step 3b.0:
  - (if necessary)



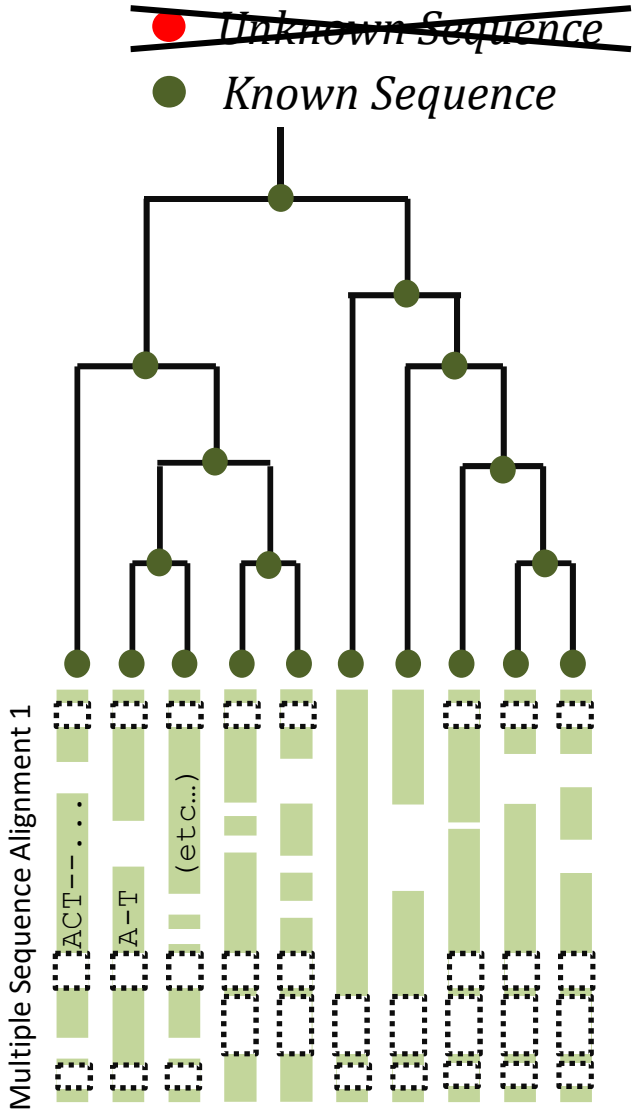

# BAli-Phy: Algorithm

- Step 3(…):
  - Choose an internal branch to start with.
  - Re-sample its branch length, conditional on the sequences on either side of it.
- Step 3(…).0:
  - (if necessary…slightly different treatment because we are at a leaf node.)




# BAli-Phy: Algorithm

- Step 4:
  - Write down the alignment and tree at this point in our list of "samples"
  - This is one our sample points for the posterior distribution of the alignment-tree combination.
- Go to Step 2a.
  - Repeat ad infinitum.



# BAli-Phy: After It Has Finished

- All this sampling gives us a huge list of alignments and tree topologies, which we give equal "weight".
  - At each point, we also have the prior probability and likelihood of the alignment, and several other parameters.
  - One way of getting a "best" estimate of the alignment or tree is to take the one with the highest (Prior) x (Likelihood) score. This is called the *Maximum A Posteriori* (MAP) alignment/tree.
  - For the alignment, another way is to calculate an alignment whose columns *collectively* appear the most often in the sample list. (Called the *Posterior Decoding* alignment.)

# Advantages & Disadvantages

**Advantages:**

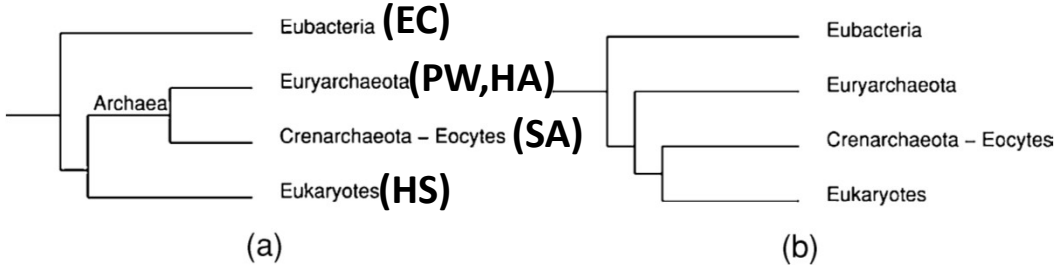
- Doesn't rely on a particular input tree or alignment.
- Does not assume indels are missing data, so tree & alignment more likely to agree.
- Tests seem to show it works pretty well for alignments, at least on small enough data.
- Outputs a distribution, not just a single estimate.
- Nice theoretical properties under the assumptions of the model.

**Disadvantages:**

- Slow, even when parallelized.
- Requires a "burn-in" period before the samples can truly have equal weight.
- MAP alignment and tree are unstable and authors don't recommend them.
- Results require interpretation (i.e. point estimate not automatically generated.)
- "Effective Sample Size" is considered important by statistical community.

# Evaluation: Example from Paper

- Authors show that fixing an alignment that is based on a guide tree (e.g. Clustalω) can overestimate the confidence in subsequent estimation of the tree.
  - Question: Do the Archea form a monophyletic group (a) or not (b)? From the paper:




  - Answer: That depends on whether you use a Clustal alignment:

| | Posterior Probability According to: | | |
|---|---|---|---|
| **Tree** | **BAli-Phy** | **BAli-Phy, fixed Clustal alnmt.** | **Other Bayesian Method, Clustal alnmt.** |
| **((EC,HS),(SA,(PW,HA)))** | 0.308 | 0.996 | 0.172 |
| **((EC,HS),((SA,PW),HA))** | 0.208 | 0.002 | 0.7 |
| **(EC,((HS,SA),(PW,HA)))** | 0.120 | 0.001 | <0.001 |
| **((EC,PW),((SA,HS),HA))** | 0.088 | <0.001 | <0.001 |
| **((EC,SA),(HS,(PW,HA)))** | 0.066 | 0.001 | <0.001 |
| **((EC,HS),(PW,(SA,HA)))** | 0.037 | <0.001 | 0.127 |
| **EC,HS\| HA, PW, SA** | 0.553 | 0.998 | >0.999 |
| **EC,HS,SA \| PW, HA** | 0.494 | 0.998 | 0.173 |

# Evaluation: Example from Paper

- By accounting for weakly supported homologies in the alignment, a better estimate of the tree can emerge.
- The image below (from the paper) shows an Alignment Uncertainty plot from the paper, displaying areas where the alignment is uncertain (light squares) and relatively certain (dark). This alignment adds credibility to tree (b) in the previous slide.

# Conclusion

- BAli-Phy gives good alignments for reasonably small numbers of taxa, and helps estimate the relative likelihood of various tree topologies.
  - Quality of estimation of topology is more dependent on dense sampling.
- Restricted to small data (e.g. < 25 taxa), and must run for >24 hours, possibly more.
- Good especially when both alignment and tree are uncertain.

**Questions?**