Calculating a likelihood on a tree

Let X refer to the data – a set of characters. T denotes the tree and $\boldsymbol{\nu}$ is the set of branch lengths. The maximum likelihood tree, $\hat{T}$, is the tree that allows us to achieve the highest value of the likelihood:

$$\mathbb{P}(X \mid T, \boldsymbol{\nu}) \tag{1}$$

If there are M characters, and we assume that the characters evolve independently of each other, then we can calculate the likelihood as a product over all characters. If x_i denotes character i, then

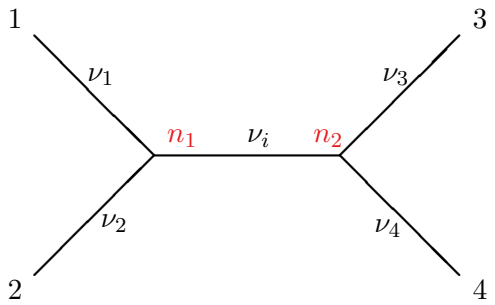
$$\mathbb{P}(X \mid T, \boldsymbol{\nu}) \quad = \quad \prod_{i=1}^{M} \mathbb{P}(x_i \mid T, \boldsymbol{\nu}) \tag{2}$$

where | is read as "given that." This means the product (over all characters) of the probability of a character identical to x_i being generated on tree T with branch lengths, $\boldsymbol{\nu}$.
Unfortunately, we do not know the ancestral states, and it is not obvious how we should assign probabilities for different scenarios of mutations/substitutions.
Imagine that we want to calculate the likelihood for the matrix:

	Characters	
Taxa	1	2
1	A	C
2	A	C
3	C	C
4	A	C

on the tree in figure (1) (with n_1 and n_2 denoting the internal nodes).

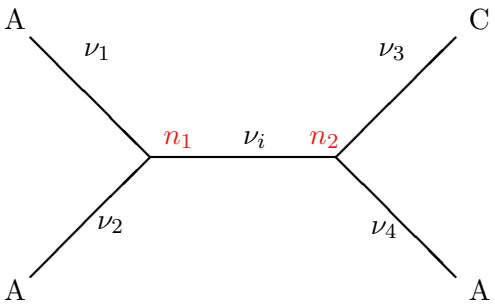
Figure 1: tree to be scored



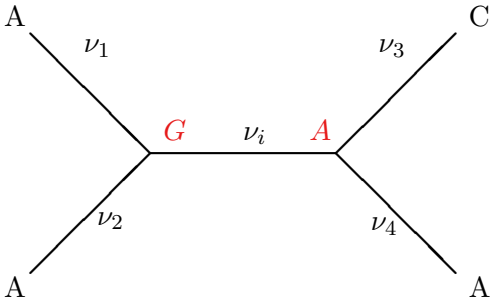
Fortunately, there are a finite number of possible ancestral character assignments. Since each possibility is mutually exclusive, we can add the probabilities for each of the possible assignments:

$$\mathbb{P}(x_i \mid T, \boldsymbol{\nu}) = \sum_{n_1 \in \{A,C,G,T\}} \sum_{n_2 \in \{A,C,G,T\}} \mathbb{P}(x_i, n_1, n_2 \mid T, \boldsymbol{\nu}) \tag{3}$$

It turns out that we often calculate likelihoods using "time-reversible" models this means that the likelihood is the same regardless of where we root the tree. So to score the first character, we can arbitrarily root the tree at anywhere. For this example, we'll root the tree at node n_1.
So if we label the tips of the tree with character #1, we get:



To do the summation in equation 3 we need to be able to calculate the probability of a set of tip states To calculate the probability of a particular pattern, we can assume that the process of evolution acts independently from character to character.
For a specific assignment of character states this means that we have problem of calculating the probability of:



given the tree and branch lengths. The assumption of independence lets us break this down into the probability of the root being G multiplied by the probabilities associated with each edge of the tree:

$$\begin{aligned}\mathbb{P}(x_i, n_1 = G, n_2 = A \mid T, \boldsymbol{\nu}) \quad = \quad & \mathbb{P}(n_1 = G) \times \mathbb{P}(A \text{ at } 1 \mid \nu_1, n_1 = G) \times \mathbb{P}(A \text{ at } 2 \mid \nu_2, n_1 = G) \\ & \times \mathbb{P}(n_2 = A \mid \nu_i, n_1 = G) \times \mathbb{P}(C \text{ at } 3 \mid \nu_3, n_2 = A) \times \mathbb{P}(A \text{ at } 4 \mid \nu_4, n_2 = A)\end{aligned}$$

So this means that we can calculate the probability of a pattern arising on a tree if we can decide how to describe

1. the probability of a state at the root, and
2. the probability of an end-state conditional on the start state and the branch length.