1 Introduction

Last week we discussed single branch. In this lecture we’ll evolve to the whole tree. The tree holds information regarding ancestors, on which we’d like to ask various questions. One of the alternatives is to count the number of changes and to pick the most reasonable ancestors - the one that requires the least amount of changes, which is very different from maximum likelihood since it ignores the branches lengths (as we did previous in Parsimony).

**Probabilistic Model** Each site is regarded separately, we define a random variable which defines what happened in each vertex. It is their distribution we are looking for, under the IID assumption. We’ll examine the root (soon we’ll assume deriving from the stationary distribution) from which a split occurred, the assumption is: Given the root the occurrence in every splitting branch is independent, to be put differently - the sons are independent of each other, given the father. Most changes being neutral makes it easy for us to assume this.

![Figure 1: A Rooted Tree](image)
Following equation is due to independent assumption, yet another such assumption we make is that given a direct ancestor, an indirect ancestor is irrelevant - the Markov assumption.

So far we made two assumptions: the speciation independence and the Markov assumption both results in:

\[ P(x_1 \ldots x_n) = P(x_n) P(x_1 \ldots x_{n-1} | x_n) = P(x_n) P(\{x_i | i \in L\} | x_n) P(\{x_i | i \in R\} | x_n) \]

Recursively doing so we end up with:

\[ P(x_1 \ldots x_n) = P(x_{\text{root}}) \prod_{i \neq \text{root}} P(x_i | x_{a(i)}, t_{i,a(i)}) \]

where \( a(i) \) is the direct ancestor of node \( i \), and \( t_{i,a(i)} \) is the time to reach \( i \) from \( a(i) \). In fact what we have is a split Markov chain, with a slight change such that each branch has a distinguished length.

**Consequences of Markov** The analogy to Markov chain raise few questions and one observation:

1. What is the joint distribution \( P(\{\text{leafs}\}) \)
2. Given evolution observation what is the probability for a given root.
3. How do we evaluate the branches lengths - the tree topology.
4. Notice that extinct species taken out side of the sum result in a sum of one multiplying all the rest - meaning have no significance. Eventually we have a tree trimmed from extinct species. (was given as an exercise)

![Figure 2: A Rooted Tree](image)

What is the probability of the observations?

\[ P(x_1, \ldots, x_n) = \sum_{x_{k+1}} \ldots \sum_{x_n} P(x_1, \ldots, x_n) \]

Where \( x_1, \ldots, x_k \) are the observation and \( x_{k+1}, \ldots, x_n \) are junctions within the tree. There are \( k - 1 \) vertices in the tree and \( 4^{k-1} \) possible assignments.
The tree structure and the two hypothesis we’ve made yield dynamic programming. As in HMM we will develop a Backward function - given the position of a hidden state what is the probability of the chronically following states.

\[
up[i, a] = P(\pi_{S(i)} | x_i = a)
\]

Where \(\pi_{S(i)}\) is all of the observations (leaves) in the subtree rooted at node \(i\)

Our end case, when \(i\) is a leaf is:

\[
up[i, a] = \begin{cases} 
1 & a = x_i \\ 
0 & \text{otherwise} 
\end{cases} 
\]

\[
* = \sum_b P(x_{S(j)}, x_j = b | x_i = a) = \sum_b P(x_{S(j)} | x_i = a) P(x_j = b | x_i = a) = \sum_b up[j, b] P(t_{i,j})_{a,b}
\]

Where \(P(t_{i,j})_{a,b}\) is a value from the Rate Matrix. We get:

\[
up[i, a] = \left[ \sum_b up[j, b] P(t_{i,j})_{a,b} \right] \left[ \sum_c up[k, c] P(t_{i,k})_{a,c} \right]
\]

Back to our question - the marginal probability over all observations:

\[
P(x_{S(root)}) = \sum_a up[root, a] P(x_{root} = a)
\]

Since \(x_{S(root)}\) and all the observations.

Our second question:

\[
P(x_{root} | x_{S(root)}) \propto \text{bayes} P(x_{S(r)} | x_r) P(x_r)
\]
**HMM extension**  What about inner vertices? Notice that we learn about every inner vertex from all of the observations, not just from its direct descendants, and is solvable using up functions. We will plan a Down function:

$$P(x_i = a | x_{S(i)}, x_{\overline{S(i)}}) = \frac{P(x_i = a, x_{S(i)}, x_{\overline{S(i)}})}{P(x_{S(i)}, x_{\overline{S(i)}})}$$

For the simplicity we will observe:

$$P(x_i, x_{S(i)}, x_{\overline{S(i)}}) = P(x_{S(i)} | x_i, x_{\overline{S(i)}}) P(x_i, x_{\overline{S(i)}}) =$$

Applying the Markov assumption:

$$= \underbrace{P(x_{S(i)} | x_i)}_{up} \underbrace{P(x_i, x_{\overline{S(i)}})}_{down} = up[i, a] down[i, a]$$

Which is what the complementary tree indicates through its ancestor. The end case would be:

$$down[root, a] = P(x_{root} = a) = \prod_{a}$$

Since $x_{root} = \varnothing$
Our assumptions helped us make the expression converge to:

$$down[j, a] = \sum_b down[i, b] P(t_{i,j})_{b,a} \sum_c up[k, c] P(t_{i,k})_{b,c}$$

And we end up with an HMM being generalized to the form of a tree.

Notice:

1. You can not go down before you went up
2. In means of calculation, we are linear in number of vertices, and quadratic in $\sum$
3. caching the up calculations using linear programming is plausible

We can learn about the inner vertices given a tree topology and observations. Another question which arise: what is the joint probability of the ancestors given the observations of the ancestors. Obviously given leaves the number of possibilities to ancestors is exponential, but we can alter the question to different one: what is most probable for all the ancestors at once to be.

1. $\arg \max_{x_i} P(x_i|x_{S(root)})$
2. $\arg \max_{x_1, \ldots, x_k} P(x_1, \ldots, x_k|x_{S(root)})$ a very difficult question not resembling choosing a value to every $x_i$ independently in the most probable way for the tree - which is up like we did but with $\max$ instead of $\sum$, like in Viterby.
The question of $P(x_i | x_{S(root)})$ we already answered $(up\cdot down)$

The next question:

$$P(x_i = a, x_{a(i)} = b | x_{S(root)})$$

Is interesting since it allows us to evaluate the branch length again and like in HMM we will get (after altering the question from a dependant probability to joint one):

$$P(x_i = a, x_{a(i)} = b | x_{S(root)}) = up[i, a]down[a(i), b]P(t_{i,a(i)}b,a)\sum_c up[j, c]P(t_{a(i),j}b,c)$$

Now we acquired the tools to do EM!

### 2 Learning Branch Lengths - EM model

Given the rate matrix, tree topology, and the observations the final statistics are looking at the whole sequence how many times did $i$ changed to $b$. The tree we are looking at is for the whole sequence.

$$N_{i\rightarrow j}[a, b] = \sum_m 1\{x_i[m] = a, x_j[m] = b\}$$

Meaning that on column $m$ the vertex $i$ changes to vertex $j$ from $a$ to $b$

$$\mathbb{E}[N_{i\rightarrow j}[a, b]|D, T, \vec{t}]_{E\ \text{step}} = \sum_m P_T(x_i[m] = a, x_j[m] = b | x_{S(root)}[m])$$

Our EM steps are:

**E-Step:** given a tree $T$ we’ll guess $\vec{t}$ the branches lengths and than up-down for each position and $\mathbb{E}[N_{i\rightarrow j}]$ (which is a matrix for every $a$ and $b$) for each branch.

**M-Step:** maximize $t_{i,j}$ based on $\mathbb{E}[N_{i\rightarrow j}]$ formally:

$$t_{i,j}^{new} \leftarrow \arg \max_t \sum_{a} \sum_{b} \mathbb{E}[N_{i\rightarrow j}[a, b]] \log P(t)_{a,b}$$

The process will converge almost always to a global maximum since the parameter space is very poor (branch length). Both the topology and the shifts are fixed. This in contradiction with HMM where the parameter space is huge. A global quality of EM, is observed also here - very slow progress at the end of the running of the algorithm. This described model is the simplest likelihood model of a protein evolution. We totally avoiding learning the tree topology. The model assumption is same evolution rate for each column independently - the simplest assumption. This assumption raises a problem: the evolutionary stress is not equal on all of the proteins’ positions (active sites, etc.) A 100% conserved position will result in extremely short length - yielding less changes. Less conserved positions yield messy tree, a lot of combinations in significant weights. Such extreme positions tell us something that our model miss. The rate of change tell us about the relevant time scale, the slower the changes, the closer we are to the root. We can enrich our model to support variating evolutionary rates.

### 3 Rate Variation Model

$r$ - rate - for every position, defining the scale from each and every position, a multiplicative factor, which changes the branch length.

$$P(x_{S(root)}[m], T, \vec{t}, r[m])$$
The assumption is that each position has a different rate. However, introducing a parameter for position makes the model much more complex.

\[
P(x_{S(root)}) = \int_{r=0}^{\infty} \sum_{x_{k+1}, \ldots, x_n \text{ancestors}} P(x_1, \ldots, x_n | \vec{t}, r) \, P(r) \, dr
\]

\[
\text{same probability for every position}
\]

\[
\text{different preservation rate} = \text{different probabilities}
\]

We are independent in scaling the factor for all of the branches in the same proportion since the likelihood remains. Due to this fact we can normalize \( \vec{t} \) and \( P(r) \) so \( E[r] = 1 \). The \( r \) values are non-negative.

**Gamma Distribution** - the positive equivalent of the Gaussian. A complete Gamma function is:

\[
\Gamma(x) = \int_0^1 [\ln \frac{1}{t}]^{x-1} dt = \int_0^\infty t^{x-1} e^{-t} dt
\]

And is considered as an extension of factorial since for every natural \( n \) you get that \( \Gamma(n) = n! \). Now we are ready to present the distribution which arises naturally in processes which the waiting times between Poisson distributed events are relevant. The standard (expectation of zero and variance of one) gamma distribution is:

\[
P(x|\alpha) = \frac{x^{a-1}e^{-x}}{\Gamma(\alpha)}
\]

The integral is being analyzed using numeric analysis such as:

\[
\int f(x)dr \sim \sum_{i=1}^{k} \omega_i f(r_i)
\]

Choosing the sampling points where we measure \( \omega \) is a question out of the scope of this paper. Another numeric method:

\[
\int g(r) P(r)dr \sim \sum_{i=1}^{k} v_i g(r_i)
\]

Where we can use some prior knowledge of the distribution. The integral turn to be rather simple ...

**Miscellaneous** Our problem transforms to calculating likelihood of few tree variants. This is the first model to have done that, and have yielded significant differences.

Other aspects of variating evolutionary stress: compensatory mutations, have no good model.

Another thing that people tend to do is to state that one position preservation inference the preservation of its neighbors and the variation rate changes locally through the protein. Structural priors such as helixes and loops can also be of assistance in dimension reduction of the problem. Meaning that a-priori structural information regarding a protein can be internalized as a prior for the rate matrix for instance. Using solved structure one can learn different rate matrix for every structure. The stationary distribution can be changed in accordance to the structure.

Another aspect is an evolutionary gain of function on a certain branch which is expressed in a change of length through some sub-tree, this can be added to the model by adding a parameter, for example.