COMP 571: Homework #4

Vi Nguyen (vkn1)

April 18, 2017

On my honor, I have neither given nor received any unauthorized aid on this homework.

1. We want to compute the maximum likelihood edge lengths t_1 and t_2. However, because there can only exist one edge between the two leaves x^1 and x^2, so the edge between them will have length $t_1 + t_2$. Let $t = t_1 + t_2$. Thus, let us find the value for t that will maximize likelihood, since there is only one possible tree topology.

   The likelihood function L for this tree is

   $$L = (P_{i\to i}(t))^{n_1}(P_{i\to j}(t))^{n_2}$$

   Finding t that maximizes L is equivalent to finding the t that maximizes the log-likelihood $\log(L)$.

   $$\log(L) = n_1 \log(P_{i\to i}(t)) + n_2 \log(P_{i\to j}(t))$$

   We can maximize the log-likelihood by taking the partial differential of $\log(L)$ with respect to t and finding where the differential equals 0, given that $P_{i\to i}(t) = \frac{1}{4}(1 + \frac{1}{3}e^{-4\alpha t})$ and $P_{i\to j}(t) = \frac{1}{4}(1 - e^{-4\alpha t})$.

   $$\frac{\partial \log(L)}{\partial t} = n_1 \frac{\partial}{\partial t} \log(P_{i\to i}(t)) + n_2 \frac{\partial}{\partial t} \log(P_{i\to j}(t)) = 0$$

   $$\Rightarrow n_1 \frac{\partial}{\partial t} \log(\frac{1}{4}(1 + \frac{1}{3}e^{-4\alpha t})) + n_2 \frac{\partial}{\partial t} \log(\frac{1}{4}(1 - e^{-4\alpha t})) = 0$$

   $$\Rightarrow \frac{n_1}{\frac{1}{4}(1 + \frac{1}{3}e^{-4\alpha t})} + \frac{n_2}{\frac{1}{4}(1 - e^{-4\alpha t})} = 0$$

   $$\Rightarrow \frac{n_1(\frac{1}{4}(1 - e^{-4\alpha t})) + n_2(\frac{1}{4}(1 + \frac{1}{3}e^{-4\alpha t}))}{\frac{1}{4}(1 - e^{-4\alpha t})\frac{1}{4}(1 + \frac{1}{3}e^{-4\alpha t})} = 0$$

$$\Rightarrow n_1(\frac{1}{4}(1-e^{-4\alpha t})) + n_2(\frac{1}{4}(1+\frac{1}{3}e^{-4\alpha t})) = 0$$

$$\Rightarrow n_1(1-e^{-4\alpha t}) + n_2(1+\frac{1}{3}e^{-4\alpha t}) = n_1 - n_1e^{-4\alpha t} + n_2 + n_2\frac{1}{3}e^{-4\alpha t} = 0$$

$$\Rightarrow n_1 + n_2 = n_1e^{-4\alpha t} - n_2\frac{1}{3}e^{-4\alpha t}$$

$$\Rightarrow n_1 + n_2 = e^{-4\alpha t}(n_1 - \frac{n_2}{3}) \Rightarrow \frac{n_1+n_2}{n_1-\frac{n_2}{3}} = e^{-4\alpha t}$$

$$\Rightarrow \frac{3(n_1+n_2)}{3n_1-n_2} = e^{-4\alpha t} \Rightarrow \ln(\frac{3(n_1+n_2)}{3n_1-n_2}) = -4\alpha t$$

$$\Rightarrow t = t_1 + t_2 = \frac{1}{4\alpha}\ln\frac{3(n_1+n_2)}{3n_1-n_2}$$

2. **Bottom-Up Phase**

   For each node v and character c compute the set $S_{c,v}$ as follows:

   - If v is a leaf, then $S_{c,v} = \{v_c\}$
   - If v is an internal node and has n children $w_1, ..., w_n$:

   $$S_{c,v} = \begin{cases} S_{c,w_1} \cap ... \cap S_{c,w_n} & S_{c,w_1} \cap ... \cap S_{c,w_n} \neq \emptyset \\ S_{c,w_1} \cup ... \cup S_{c,w_n} & otherwise \end{cases}$$

   Top-Down Phase

   - For the root r, let $r_c = a$ for some arbitrary a in set $S_{c,r}$
   - For internal node v with parent u and α are all of the elements in $S_{c,v}$

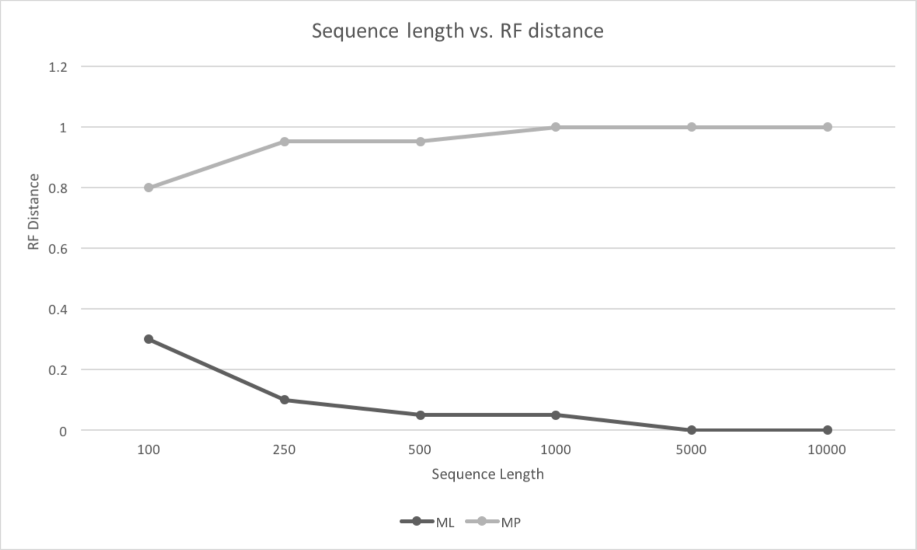
   $$v_c = \begin{cases} u_c & u_c \in S_{c,v} \\ argmin_\alpha(W(u_c,\alpha)) & otherwise \end{cases}$$

3. We can see from Fig. 1 that as sequence length of generated sequences increased, the accuracy (based on RF distance) of trees generated using maximum likelihood increased while trees generated using maximum parsimony decreased.

   However, it seems logical that both methods of determining evolutionary history would become more accurate as sequence length increased, since more data is available.

   The increasing error in MP could be due to the fact that when running MP, there was not an option to take into account the base frequencies of nucleotides, while there was for ML. This could have skewed the results as sequence lengths became longer. In addition, it is possible that as sequence length increased, the amount of noise and "false positives" for sites that appear to be related to each other, which led MP to construct trees dissimilar to the original one.

Figure 1: Chart for problem 3.



4.