Algorithmic Bioinformatics DD2450, spring 2010, Lecture 13

Lecturer Jens Lagergren
Several current and previous students
will be acknowledged in a separate document.

May 22, 2010

1 Bayesiansk inferens in Phylogeny

1.1 Significance Measure for a Tree

In Bayesian inference we want to compute the posterior probability (the conditional probability that is assigned after the relevant evidence is taken into account), which is out measure for significance. Let $s_1, \ldots, s_n$ be our observed sequences and T be a tree, we want to compute the posterior

$$\Pr[T|s_1, \ldots, s_n]$$

for the trees T with the "highest posterior".
For trees with edge length λ

$$\Pr[T|s_1, \ldots, s_n] = \int_\lambda \Pr[T, \lambda|s_1, \ldots, s_n]$$

and

$$\Pr[T, \lambda|s_1, \ldots, s_n] = \frac{\Pr[s_1, \ldots, s_n|T, \lambda]\Pr[T, \lambda]}{\Pr[s_1, \ldots, s_n]}$$

where

- $\Pr[T, \lambda]$ is a prior, that is to say the probability of the weighted tree T before we did any observation. This probability is given in whole by our model. We will use a uniform distribution for the posterior but there are cases where other distributions are better but the uniform case is easier to work with.
- $\Pr[s_1, \ldots, s_n]$ can be written as $\sum_T \int_{\lambda_T} \Pr[s_1, \ldots, s_n|T, \lambda]$ but this is hard to compute. In general we can't compute probabilities but we *can compute ratios between probabilities. In this case we want to compute the fractions of two posteriors.*

$$\frac{\Pr[T,\lambda|s_1,\ldots,s_n]}{\Pr[T',\lambda'|s_1,\ldots,s_n]} = \frac{\frac{\Pr[s_1,\ldots,s_n|T,\lambda]\Pr[T,\lambda]}{\Pr[s_1,\ldots,s_n]}}{\frac{\Pr[s_1,\ldots,s_n|T',\lambda']\Pr[T',\lambda']}{\Pr[s_1,\ldots,s_n]}} = \frac{\Pr[s_1,\ldots,s_n|T,\lambda]}{\Pr[s_1,\ldots,s_n|T',\lambda']}.$$

and for this we can use MCMC.

- First we create a Markov Model M with states $\{T,\lambda\}$. We set the transition probabilities of M so that M's stationary distribution becomes the posterior. $\varphi(T,\lambda) = \Pr[T,\lambda|s_1,\ldots,s_n]$.
- Then we estimate $\varphi(T,\lambda)$ by traversing M.
  - first we do a burn-in (i.e. we traverse M for a long time - something in the range of 100'000 transitions).
  - second we sample states at defined intervals. (Something in the range of every 100 step).
- We now want to compute $\int_\lambda \varphi(T,\lambda)$ (for various T) and we estimate it by the fraction of samples with T together with some λ.

In MCMC transition probabilities are determined by two distributions.

1. A *proposal distribution* , i.e. in state $\{T,\lambda\}$ another state $\{T',\lambda'\}$ is proposed with probability $\rho(T',\lambda'|T,\lambda)$.
2. A *acceptance distribution* , i.e. in state $\{T,\lambda\}$ another state $\{T',\lambda'\}$ is proposed. We accepted the new state with probability $\alpha(T',\lambda'|T,\lambda)$ (if the new state isn't accepted then was stay in $\{T,\lambda\}$).

The transition probability is given by the compose of the proposal probability and the acceptance probability.

$$\mathcal{T}(T',\lambda'|T,\lambda) = \rho(T',\lambda'|T,\lambda)\alpha(T',\lambda'|T,\lambda)$$

In MCMC, no acceptance probability is 1, so each state has a loop.

Definition 1. *A Markov Model is*

1. irreducible *if for each pair of states* (a,b) *the probability of going from* $a \to b$ *in one or more steps is* > 0
2. periodic *if for a state a* $Pr[a \to a \text{in } i \text{ steps}] > 0 \Longrightarrow k|i$

Sats 1. *Let M be a irreducible, aperiodic Markov-chain with proposal distribution* ρ *and acceptance distribution* α*. Then if*

1. ρ *is symmetric* ($\rho(a|b) = \rho(b|a)$).
2. $\alpha(a|b) = \min(1, \frac{\varphi(a)}{\varphi(b)})$

then φ is it's stationary distribution.

Proof. The transition probabilities are given by $\mathcal{T}(T', \lambda'|T, \lambda) = \rho(T', \lambda'|T, \lambda)\alpha(T', \lambda'|T, \lambda)$. Notice that the stationary distribution is the unique distribution that satisfies the detailed balance equation.

$$f(a) = \mathcal{T}(b|a) = f(b)\mathcal{T}(a|b)$$

Let a, b be two arbitrary states. *Assume* $\varphi(a) \geq \varphi(b)$ hence $\alpha(b|a) = \frac{\varphi(a)}{\varphi(b)}$ and $\alpha(a|b) = 1$.

$$\begin{aligned}\varphi(a)\mathcal{T}(b|a) &= \varphi(a)\rho(b|a)\frac{\varphi(b)}{\varphi(a)} \\ &= \varphi(b)\rho(a|b) \\ &= \varphi(b)\varphi(a)\mathcal{T}(a|b)\end{aligned}$$

so φ satisfies the detailed balance equation.

■

1.2 In Practice

There is one question we want to answer in practice. What is the proposed distribution? (This is all we need to do MCMC in phylogeny). so we Pick an edge e uniformly from $E(T)$.

- With probability $\frac{1}{11}$ stay.
- With probability $\frac{5}{11}$ do a branch swap on e.
- With probability $\frac{5}{11}$ pick a number s from a normal distribution with mean 0 if $\lambda(e) + s < 0$ stay otherwise let $\lambda(e) = \lambda(e) + s$ $\lambda(e)$ with 2.