

ABC ESTIMATION OF A SUBSTITUTION RATE.  
GEOFF NICHOLLS, DTC 24/03/10

Refer to <http://www.stats.ox.ac.uk/~nicholls/dtc/TT10/> for material.

We will begin with a practical on ABC estimation of the parameter of a Poisson distribution from a few independent realizations of a Poisson distributed random variable. We will then move on to the two main exercises. The first of these exercises has two more or less independent parts, and you can choose which to do.

**9:30:** ABC

**10:30:** Practical on ABC

**11:30:** Tree models and substitution models

**14:30:** ABC for Bayesian inference

**15:00:** Practical on ABC for substitution rate from sequence data.

SUBSTITUTION MODEL

Let  $x$  and  $y$  be two aligned sequences of AA bases,  $x = (x_1, x_2, \dots, x_N)$  etc, with  $x_i, y_i \in \{A, C, G, T\}$  say.

Suppose  $y$  descends from  $x$  over a time interval of length  $\delta$ . In the neutral independent finite sites substitution model, events at each site are independent, we assume neutral evolution, and ignore processes such as indels which change the sequence length.

Since the sequences are aligned, the base at site  $i$  in  $y$  evolves by substitution from the base at site  $i$  in  $x$ . Let

$$P_{a,b} = \Pr(y_i(\delta) = b | x_i(0) = a)$$

and

$$\pi_a = \Pr(x_i = a).$$

Here  $P_{a,b}$  models the probability for different kinds of changes, as a function of time, and  $\pi_a$  gives the equilibrium proportions of the different bases (since we arrive at sequence  $x$  following simulation from the infinite past). It is not too hard to show that

$$\pi_b = \sum_a \pi_a P_{a,b}$$

for the equilibrium base frequency vector  $\pi$ , or in matrix notation  $\pi = \pi P$ .

**Parameterization, and the rate Matrix,  $Q$ .** We will parameterize  $P$  via the rates  $Q_{a,b}$  for  $a \rightarrow b$  transitions. Clearly  $P_{a,b} = P_{a,b}(t)$ . We can see that  $P_{a,b}(0) = 0$  if  $a \neq b$  and  $P_{a,b}(0) = 1$  if  $a = b$ , so  $P(0) = I_4$  where  $I_4$  is the  $4 \times 4$  identity matrix. Parameterize

$$\begin{aligned} P_{a,b}(\delta) &= Q_{a,b}\delta + o(\delta) & b \neq a \\ P_{a,a} &= 1 + Q_{a,a}\delta + o(\delta) \\ P &= I_4 + Q\delta + o(\delta), \end{aligned}$$

the last line in matrix notation. The little- $o$  notation is  $f(\delta)$  is  $o(\delta)$  if  $f(\delta)/\delta \rightarrow 0$  as  $\delta \rightarrow 0$ . These equations are just the Taylor expansion of  $P$ , using  $P(0) = I_4$ . In order to get  $\sum_b P_{a,b} = 1$  at  $o(\delta)$  we need  $Q_{aa} = -\sum_{b \neq a} Q_{a,b}$ . The  $Q_{a,b}$  ( $a \neq b$ )

parameters are important physical parameters, modeling the instantaneous rate at which an  $a$  is substituted by a  $b$ . However the diagonal elements of  $Q$  are determined from the others elements - we do not wish to model  $Q_{a,a}$ , the rate at which nothing happens! In order to fix the large  $\delta$  behavior of  $P(\delta)$  we impose a homogeneous Markov structure: The future is independent of the past given the present. For  $\tau \in (0, t)$ ,

$$P_{a,b}(t) = \sum_{c \in \{A,C,G,T\}} P_{a,c}(\tau) P_{c,b}(t - \tau),$$

or in matrix notation  $P_{a,b}(t) = [P(\tau)P(t-\tau)]_{a,b}$  (multiply the matrices and take the  $(a, b)$  element). We can use this idea to cut a macro-time interval into micro-time pieces, in which our defining equations for  $P(\delta)$  hold good.

$$\begin{aligned} P(t) &= P^2(t/2) \\ &= P^n(t/n) \\ &= (I_4 + Qt/n + o(t/n))^n \\ &\rightarrow e^{Qt}, \end{aligned}$$

where the last line is obtained by taking the limit as  $n \rightarrow \infty$  and using  $(I + A\delta)^{1/\delta} \rightarrow e^A$  where  $A$  is a suitable matrix (meaning, one for which  $e^A = I + A + A^2/2 + \dots$  is defined).

This last expression  $P(t) = \exp(Qt)$  is a matrix exponential. We could compute it by diagonalizing  $Q = UDU^T$  with  $D$  a diagonal matrix of eigenvalues, and writing  $P(t) = U \exp(Dt)U^T$ . Here  $\exp(Dt)$  is a diagonal matrix with entries  $\exp(Dt)_{ii} = \exp(D_{ii}t)$ . We can compute matrix exponentials in MatLab using `expm()`.

**Units of time.** You can check from the relation between  $P$  and  $Q$  that if  $\pi$  is the EBF for  $P$  so  $\pi = \pi P$  then  $\pi Q = 0$ . We usually scale  $Q$  to have units substitutions, so there is one substitution per unit time.  $Q$  generates  $-\pi \cdot \text{diag}(Q)$  substitutions per unit time, so if this isn't equal to one we set  $Q \rightarrow -Q/\pi \cdot \text{diag}(Q)$  to get the desired units.

**Jukes Cantor Example.** One very simple substitution model is the Jukes-Cantor model. In this model all the off diagonal entries in  $Q$  are equal. It follows that  $\pi = (1, 1, 1, 1)/4$  - all bases appear in equal proportions over time at a site, or along sites. In order to get a model in units of substitutions we need

$$Q = \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{pmatrix}.$$

This is a nice case where we can actually compute  $P(t) = \exp(Qt)$  in a closed form

$$\begin{aligned} P_{a,b} &= (1 + 3e^{-4t/3})/4 \quad \text{for } a \neq b, \\ P_{a,a} &= (1 - e^{-4t/3})/4. \end{aligned}$$

We can model other types of substitution, allow for an increased rate for transitions over transversions, and so on. The model is easily extended to codon substitution models, with  $61 = 64 - 3$  rather than 4 possible states.

**Simulation, down a branch.** In order to simulate the substitution process from  $x$  to  $y$  we simulate  $x_i \sim \pi$  (eg, if  $U_i \sim U(0, 1)$  then we set  $x_i$  equal the smallest  $a$  such that  $\sum_{j=1}^a \pi_j > U_i$  - this generates  $x_i = a$  with probability  $\pi_a$ ) for each  $i = 1, 2, \dots, N$ . We then compute  $P(t) = \exp(Qt)$  using `expm(Q * t)` and simulate the transitions. For Jukes-Cantor this is simple. For each  $i = 1, 2, \dots, N$  we set  $y_i = x_i$  with probability  $P_{x_i, x_i}(t)$  and otherwise we choose one of the remaining three possibilities at random. In general, simulate  $y_i$  according to the probabilities in row  $x_i$  of  $P(t)$ ,  $y_i \sim [P_{x_i, 1}, P_{x_i, 2}, P_{x_i, 3}, P_{x_i, 4}]$ . This is the same kind of discrete variate simulation we did for  $x_i$ : we take  $V_i \sim U(0, 1)$  and set  $y_i$  equal the smallest  $b$  such that  $\sum_{j=1}^b P_{x_i, j} > V_i$ .

**Simulation, on a tree.** In order to simulate synthetic data on the leaves of a tree, we will simulate a root sequence from the equilibrium base frequencies, using the scheme we just gave for simulating  $x$ . For Jukes Cantor, with its uniform EBF, this is just a sequence of  $N$  uniform random bases. Next, for each edge of the tree we simulate the sequence at the child node given the sequence at the parent node. We do this second part using the scheme we just gave for simulating  $y$  given  $x$ .

#### THE COALESCENT AND COALESCENT SIMULATION

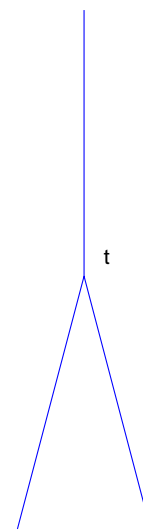
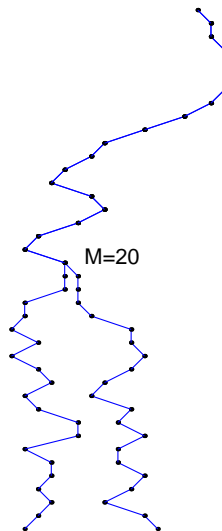
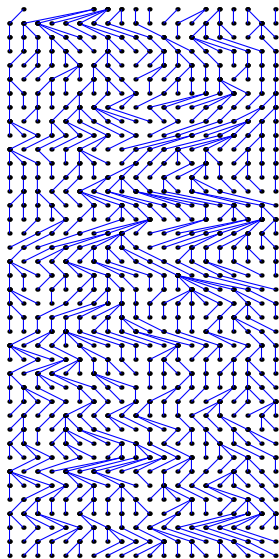
The Coalescent is a model of the genealogy of  $L$  individuals. Let time (*ie* age) increase into the past. We start out with time measured in generations.

Take a Wright Fisher population (single sex, fixed population size  $N_e$ , non-overlapping generations). In this model of ancestry each individual in the present generation chooses their parent uniformly at random from among the  $N_e$  individuals in the previous generation.

Wright-Fisher population

Wright-Fisher genealogy

Coalescent genealogy



Take two individuals at random from the present generation (*ie* without reference to their ancestry) and trace their ancestry back in time. How many generations  $M$  must we go back in time before lineages coalesce at the most recent common ancestor (MRCA)?

$$\begin{aligned}\Pr(M = m) &= \frac{1}{N_e} \left(1 - \frac{1}{N_e}\right)^{m-1} \\ \Pr(M \leq m) &= \sum_{M=1}^m \frac{1}{N_e} \left(1 - \frac{1}{N_e}\right)^{M-1} \\ &= \frac{1}{N_e} \frac{1 - \left(1 - \frac{1}{N_e}\right)^m}{1 - \left(1 - \frac{1}{N_e}\right)} \\ &= 1 - \left(1 - \frac{1}{N_e}\right)^m\end{aligned}$$

Let  $T = M/N_e$  (time in units of  $N_e$  generations). Consider what happens when  $N_e$ , the population size, is large. Using  $\lim_{a \rightarrow 0} (1 + a)^{1/a} = e$ ,

$$\begin{aligned}\Pr(T \leq t) &= 1 - \left(1 - \frac{1}{N_e}\right)^{N_e t} \\ \lim_{N_e \rightarrow \infty} \Pr(T \leq t) &= 1 - e^{-t}.\end{aligned}$$

That is, the distribution of the time back to the MRCA in the coalescent approximation to the WF model is  $T \sim \text{Exp}(1)$ . Of course real populations are not infinite, but if the limiting distribution is  $\text{Exp}(1)$  then we expect that for large populations the distribution of  $T$  will be well approximated by  $\text{Exp}(1)$ . Many different micro-time-scale models (WF, Moran,...) have this same large-time-scale limit model (coalescent).

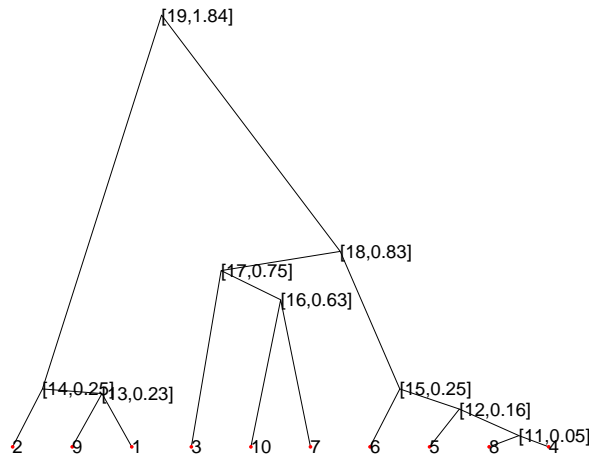
**Simulation (two leaf case).** As we move back in time, the two lineages ancestral to the two selected individuals coalesce at instantaneous rate  $\lambda = 1$ . So, to sample a two leaf tree we simply simulate  $T_1 \sim \text{Exp}(1)$ , the time back to the first coalescent event.

When we come to deal with the genealogy of  $L$  individuals we have the possibility of 3 or more lineages coalescing at the same time. Trees with events of this kind in them have a probability which is  $O(1/N_e)$  relative to trees involving only binary coalescent events (since the probability to have any particular 3-way merge is  $1/N_e^2$  compared to  $1/N_e$  for the 2-lineage case). For this reason multifurcations are not important for large  $N_e$  WF populations.

**Simulation ( $L$  leaf case).** How do we define and simulate an  $L$ -leaf coalescent genealogy?

- (1) Let  $k = L$ ,  $j = 1$ ,  $i = L + j$ ,  $t_1, t_2, \dots, t_L = 0$  and  $V = \{1, 2, \dots, L\}$ . Think of the labels in  $V$  as the labels of the leaf nodes.
- (2) As we move back in time, each pair of the  $k$  ancestral lineages we are following coalesces independently at instantaneous rate  $\lambda = 1$ . The number of pairs of lineages is  $k(k - 1)/2$  so the total instantaneous coalescent rate for  $k$  lineages is  $\lambda k(k - 1)/2 = k(k - 1)/2$ .

- (3) Simulate the length of the time interval  $T_i \sim \text{Exp}(k(k-1)/2)$  back to the next (into the past) coalescent event. Set  $t_i = t_{i-1} + T_j$  (so,  $t_{L+1} = t_L + T_1$  with  $t_L = 0$ , for the first coalescence event). The label for the tree node representing this coalescence will be  $i$ .
- (4) Conditional on there being a coalescence event at time  $t_i$ , the pair of lineages involved could be any pair with equal probability, so choose two lineages (*ie* two child nodes) at random from  $V$  and set  $V \leftarrow (V \setminus \{a, b\}) \cup \{i\}$ .
- (5) If  $i = 2L - 1$  (so should have  $k = 2$  and  $j = L - 1$ ) we are at the root so stop. Otherwise set  $j \leftarrow j + 1$  (move to the next interval),  $k \leftarrow k - 1$  (which has one less lineage) and  $i = L + j$  (the label for the next node) and go back to Step 2.



#### ABC FOR THE MUTATION RATE $\mu$

We will estimate  $\mu$  from sequence data  $D$ . Let  $S$  be the unknown true genealogy, let  $M$  be the unknown true substitution rate, and let  $D$  be an  $L \times N$  matrix of sequence data (*ie*, one  $1 \times N$  vector giving the base character at each site for each sampled individual). How did 'nature' generate the data (assume nature employs a coalescent and the Jukes Cantor substitution model)?

First  $S \sim \text{Coalescent}$ , so the genealogy was created. The sequence evolution is neutral by assumption, so the sequences then evolve down this tree without affecting it. The genealogy just provides the railway tracks for the sequence evolution. So,  $D \sim \text{Jukes Cantor}$  given  $S$ . Let  $x^{(n)}$  be the character sequence at node  $n = 1, 2, \dots, 2L - 1$ . We simulate  $x^{(2L-1)}$  (the sequence at the root) as  $x_i^{(2L-1)} \sim \pi$ , *ie*, as a draw from the equilibrium base frequency distribution. Next for  $j, k \in \{1, 2, \dots, 2L - 1\}$  with  $k$  the parent node of  $j$  on the tree  $S$

$$x_i^{(j)} \sim P_{x_i^{(k)}}(t_k - t_j) \quad \text{for } i = 1, 2, \dots, N.$$

Here  $P = \exp(\mu Qt)$  with  $\mu$  in units of substitutions per  $N_e$  generations.  $Q$  is the JC rate matrix and the time  $t_k - t_j$  is the time elapsed between the times of nodes  $k$  and  $j$ .

Our ABC will recapitulate this observation process.

Now,  $\pi(\mu, s|D) \propto p(\mu, s)p(D|\mu, s)$  with  $p(\mu, s) = p_1(\mu)p_2(s)$ . Here  $p_1(\mu)$  is a subjective prior for  $\mu$  (see below), and  $p_2(s)$  the coalescent probability density over rooted  $L$ -leaf trees.  $p(D|\mu, s)$  is the likelihood of the data. We want to avoid computing  $p(D|\mu, s)$  so we use ABC for the simulation of  $(\mu, s) \sim \pi(\mu, s|D)$ .

ABC algorithm

Aim: draw  $X \sim \pi(x|D) \propto p(D|x)p(x)$  (approximately)

Fix  $\epsilon > 0$  and some measure of distance  $d(D', D)$  between 'data sets'.

1. draw  $\mu \sim p_1(\mu)$ ,  $s \sim p_2(s)$  and  $D' \sim p(D'|\mu, s)$ .
2. while  $d(D', D) > \epsilon$  do
  - $\mu \sim p_1(\mu)$
  - $s \sim p_2(s)$
  - $D' \sim p(D'|\mu, s)$

end

return  $X = (\mu, s)$

We can run this whole thing repeatedly to collect samples  $(\mu_1, s_1), (\mu_2, s_2), \dots, (\mu_M, s_M)$  and then plot a histogram of  $\mu_1, \mu_2, \dots, \mu_M$  to get the marginal posterior for  $\mu|D$ .

How to choose  $d(D', D)$  and  $\epsilon$ ? Let

$U(D)$  = the mean number of unique site patterns, in  $D$ , per site.

Let

$$d(D', D) = |U(D) - U(D')|.$$

The idea here is that as  $\mu$  increases, the number of unique site patterns increases, so this statistic should be sensitive to  $\mu$ . Choose  $\epsilon$  as small as possible, but large enough so that the algorithm returns samples in a reasonable amount of time. Check that results are not sensitive to the choice of  $\epsilon$ .

Choose  $p_1(\mu)$  equal to the uniform distribution on  $0 \leq \mu \leq 1$ . Why?  $E(t_{2L-1}) = 2(1 - 1/L)$  since this is a coalescent (we saw this for  $L = 2$ ) so the mean root time in the tree prior is between 1 and 2. Clearly  $\mu \geq 0$ . If  $\mu$  is large enough to make  $\mu t_{2L-1} \gtrsim 1$  then the data will be saturated with substitutions. We would usually know if we had saturated data (sequences with very ancient MRCA's). Where saturated data occurs, we cannot estimate  $\mu$  other than to show that it is 'large' - here of order one or larger.

Our ABC in slightly more detail is then:

ABC algorithm

Aim: draw  $X \sim \pi(x|D) \propto p(D|x)p(x)$  (approximately)

Fix  $\epsilon > 0$  and some measure of distance  $d(D', D)$  between 'data sets'.

0. compute  $U(D)$  for  $U = \#$  unique site patterns.
1. draw  $\mu \sim U(0, 1)$ ,  $s \sim$ coalescent and  $D' \sim p(D'|\mu, s)$ .
2. compute  $U(D')$  and  $d = |U(D') - U(D)|$ .
3. If  $d < \epsilon$  stop and return a sample  $X = (\mu, s)$ , and otherwise, goto 1.