

Continuous time molecular evolution

Thomas Mailund

September 8, 2007

For analytical analysis of molecular evolution, a continuous model is mathematically more convenient. It is not really that we believe that evolution is a continuous process—of course substitutions occur as discrete events and their fixation in a species goes through a discrete process like the Wright-Fisher model—but on the time scales we work, with very long intervals between events, the continuous process is accurate enough to prefer over the mathematically less tractable discrete time process.

The continuous time model

The major change when we shift from the discrete time process to a continuous time process is that we change from working with transition probabilities and instead start working with transition *rates*. Instead of a transition probability matrix, P , we have a transition rate matrix, Q :

$$Q = \begin{pmatrix} q_{AA} & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & q_{CC} & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & q_{GG} & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & q_{TT} \end{pmatrix}$$

where q_{XY} is the (instantaneous) rate of mutating nucleotide X to nucleotide Y . Put in another way, we expect about $h \cdot q_{XY}$ mutations from X to Y to occur in a time interval of length h , so the rate of change over time t is just the (scalar) product tQ .

Just as the cells in transition probability matrices are not completely free to vary—each row must sum to 1—the rates in a transition rate matrix are not completely free to vary. The nucleotides staying the same—the diagonal of Q —should match those mutating—the off-diagonal entries of Q . Formally, we want the rate of mutations and non-mutations to sum to zero, since otherwise we would gain or lose nucleotides over time. Thus, we have the following restriction on Q :

$$q_{ii} = - \sum_{j \neq i} q_{ij}$$

called the *steady state* restriction.

To get the probability of nucleotide X changing to Y in time t —where, remember, it can go through arbitrarily many intermediate nucleotides—we need to solve the system of linear differential equations:

$$\frac{dP(t)}{dt} = QP(t) \wedge P(0) = I \quad \Rightarrow \quad P = \exp(tQ)$$

where $\exp(tQ)$ is the matrix exponential of tQ . One way to calculate this—although not necessarily the most stable numerical method¹—is to use the eigen-decomposition: $\exp(tQ) = U\Lambda U^{-1}$ where U is the matrix of eigen vectors and Λ is the diagonal matrix

$$\Lambda = \begin{pmatrix} \exp(\lambda_1 t) & 0 & \cdots & 0 \\ 0 & \exp(\lambda_2 t) & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \exp(\lambda_n t) \end{pmatrix}$$

where the λ_i are the eigen values of Q . To compute this in R you can use:

```
mexp <- function(Q,t) {
  ## calculates matrix exponential using
  ## a simple eigen-value decomposition
  x <- eigen(Q)
  U <- x$vectors
  L <- diag(exp(t*x$values))
  return (U %*% L %*% solve(U))
}
```

Jukes-Cantor in continuous time

Any rate matrix (satisfying the steady state requirement) will do, but for the exercises we will again use the Jukes-Cantor model. For the continuous time Jukes-Cantor model, α is the rate of mutation, giving us the substitution rate matrix

$$Q = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

Exercises

Simulating the evolution of a nucleotide sequence: Write a function that, given a sequence length, n , simulates X from the stationary distribution. Then write a function that, given X , the mutation probability α , and a time length, t , evolves X for t time-units to obtain the sequence Y .

Calculating the likelihood: Write a function that takes α , X , Y and t as parameters and calculates $\text{lhd}(\alpha \mid X, Y, t)$. Write another function that calculates $\text{lhd}(t \mid X, Y, \alpha)$. Try simulating X and Y (for various α s and t s) and plot $\text{lhd}(\alpha \mid X, Y, t)$ and $\text{lhd}(t \mid X, Y, \alpha)$.

¹Calculating the matrix exponential is harder than you might think, but if you do not trust me on this, google the paper *Nineteen dubious ways to calculate the exponential of a matrix*.

Estimating parameters: The two variables (or parameters) t and α are closely linked. In the likelihood, neither ever appear alone; only their product appear. As a consequence, you will never be able to estimate either, unless you know the other. Let us just fix $\alpha = 1$ and use $\text{lhd}(t | X, Y, \alpha)$ to estimate t .

Try estimating t in two different ways: use the likelihood and `nlm()` and use the analytical solution

$$\hat{t} = -\frac{1}{4} \log\left(1 - \frac{4}{3}\hat{p}\right)$$

where \hat{p} is the fraction of nucleotides that differ between X and Y . See EG pages 486–487 for how to derive this expression (but not that EG considers the time interval $2t$ and thereby get the expression $-\frac{1}{8} \log\left(1 - \frac{4}{3}\hat{p}\right)$ instead).