Process heterogeneity across sites and trees

Fredrik Ronquist

September 26, 2005

1 Introduction

So far, we have usually assumed that the evolutionary process is homogeneous. That is, the same process operates on all sites and on all parts of the phylogenetic tree. As you can imagine, however, the evolutionary process is more complex and it is often important to capture at least some of that complexity in phylogenetic models. In this lecture, we will discuss general approaches to accommodating process heterogeneity in phylogenetic models and we will discover that some previously discussed models are specific examples of these general approaches.

2 Heterogeneity across sites

There are two fundamentally different approaches we can take to process variation across sites: either the process varies independently across sites or there is some correlation between adjacent sites. In the simplest incarnation of the first approach, we assume that the process parameters x at each site are independently drawn from some probability distribution on the possible values of x. This can be generalized to mixture models where the process at each site comes from some finite mixture of processes. These models do not capture correlation between adjacent sites. If there is such correlation and site i evolves under process p_1 then we expect neighboring sites to be evolving under the same process (p_1) with a higher probability than an average site j drawn randomly from the sequence. Spatial correlation between adjacent sites is often modeled using Hidden Markov Models (HMMs).

2.1 Independently distributed process parameters

In a typical application of an independent distribution to accommodate process variation, we would focus on a parameter x in the substitution model and assume that its value is drawn from some appropriate probability distribution. As an example, let us revisit the gamma model of rate variation across sites. Assume that r_i is the rate at site i. The model specifies that $r_i \sim \text{Gamma}(\alpha, \alpha)$; in other words, the site rates are distributed (iid) according to a Gamma distribution with shape α and mean 1.0).

We still have not covered the calculation of site likelihoods, but for now, let us assume that we can calculate them. The likelihood of a site in the alignment is more clearly expressed as the probability of the observed tip data (X_i) given the rate at the site (r_i), the shape parameter of the gamma distribution (α), the other substitution model parameters σ, and the phylogenetic tree (with branch lengths) τ; we can denote this $P(X_i|r_i, \sigma, \tau, \alpha)$. To simplify the notation we will collect all the substitution model parameters in a single vector $\theta = \{\alpha, \sigma, \tau\}$.

There can be several purposes of introducing a mixture model but many times the sole purpose is to make the model more realistic. Under such circumstances, we are typically not interested in the rate at each site. Instead, we are interested in calculating and maximizing (in maximum likelihood inference) the probability or calculating the average probability (in Bayesian inference) with respect to the phylogeny parameter or potentially some other parameters in θ. To do that we need to be able to calculate and maximize the probability of the entire alignment (the data set), $P(X|\theta)$ preferably without having to worry about the rates at each site. A reasonable approach is to integrate out the rate variation at each site (we will discover later that this is really a Bayesian solution).

Since the sites are independent, even with the introduction of the mixture model, we know that

$$P(X|\theta) = \prod_i P(X_i|\theta)$$

that is, the overall probability is just a product of the site probabilities. The site probabilities are obtained by integrating or summing out the rate distribution for each site. In the case of the gamma distribution, we know that the density function is (assuming that α is the shape parameter and β is the inverse scale parameter)

$$f_\Gamma(x) = \frac{x^{\alpha-1}\beta^\alpha e^{-\beta x}}{\Gamma(\alpha)}$$

which simplifies to

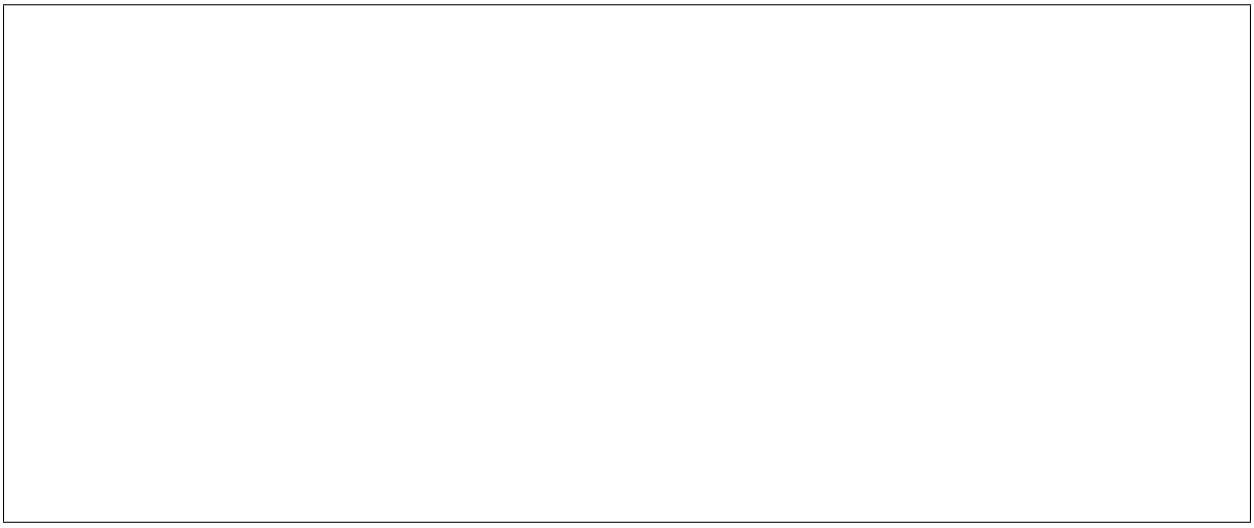
$$f_\Gamma(x) = \frac{x^{\alpha-1}\alpha^\alpha e^{-\alpha x}}{\Gamma(\alpha)}$$

when we set $\beta = \alpha$, which is the same as requiring that the mean of the distribution is 1.0 since the expectation of the gamma is α/β. Now we can calculate the site probability by integrating out the rate variation as

$$P(X_i|\theta) = \int P(X_i|r_i, \theta) f_\Gamma(r_i)\, \mathrm{d}r_i$$

with $f_\Gamma(r_i)$ expanding to the density function of the gamma distribution with mean 1.0.

Figure 1: Discrete approximation of the gamma distribution. The distribution is divided into k intervals of equal probability mass and then the rate of each category is determined as the average between the rates at the two end points.



It is not possible to calculate this integral analytically for trees of any reasonable size so it will have to be done numerically. Yang (1994) suggested a simple solution, namely to divide the gamma distribution into k discrete categories with equal probability (Fig. 1) and then calculate the integral as a sum of the site probabilities under each of these categories, using as the rate r_j for category j the mean of the two rates at the end points of the category. This numerically approximated site probability is

$$P(X_i|\theta) \approx \sum_{j=1}^{k} \frac{1}{k} P(X_i|r_j, \theta)$$

A more accurate approximation may be obtained by using Gaussian quadrature on Legendre polynomials (Felsenstein, 2001). This type of approximation uses a weighted sum, where each evaluation point r_j carries an associated weight w_j. The approximation now becomes

$$P(X_i|\theta) \approx \sum_{j=1}^{k} w_j P(X_i|r_j, \theta)$$

However, a potentially more important problem than selecting the appropriate categorization of the gamma distribution is the fit of this distribution to the actual rate variation across sites: if the fit is poor, the approximation will not be particularly accurate regardless of whether Gaussian quadrature is used.

The computational complexity of evaluating probabilities increases in direct proportion to the number of discrete rate categories used. In other words, computational considerations suggest keeping k to a minimum. Practical implementations in current phylogenetic inference programs typically default to using four categories.

The distributional approach sketched above can be used to accommodate across-site variation not only in overall rate but also in many other substitution model parameters, such as the transition/transversion rate ratio and the stationary state frequencies. The difficulty lies primarily in finding appropriate distributions of across-site variation and good numerical approximations of the site probabilities. In the Bayesian context, one can consider using MCMC to integrate out the across-site variation but there will be a large number of parameters then to integrate out (one for each site in the sequence) so it is unclear how successful this approach might be.

2.2 Mixture models

In a mixture model, the process at a site is drawn independently from k categories of process models. The different process models can have parameters that are fixed or estimated and the probability of being in each category can be estimated as well. As an illustration, consider again rate variation across sites. This time, we will assume that there are k rate categories. Each category j will have an associated rate r_j and a probability (proportion of sites) of w_j. We now calculate the site probability exactly as

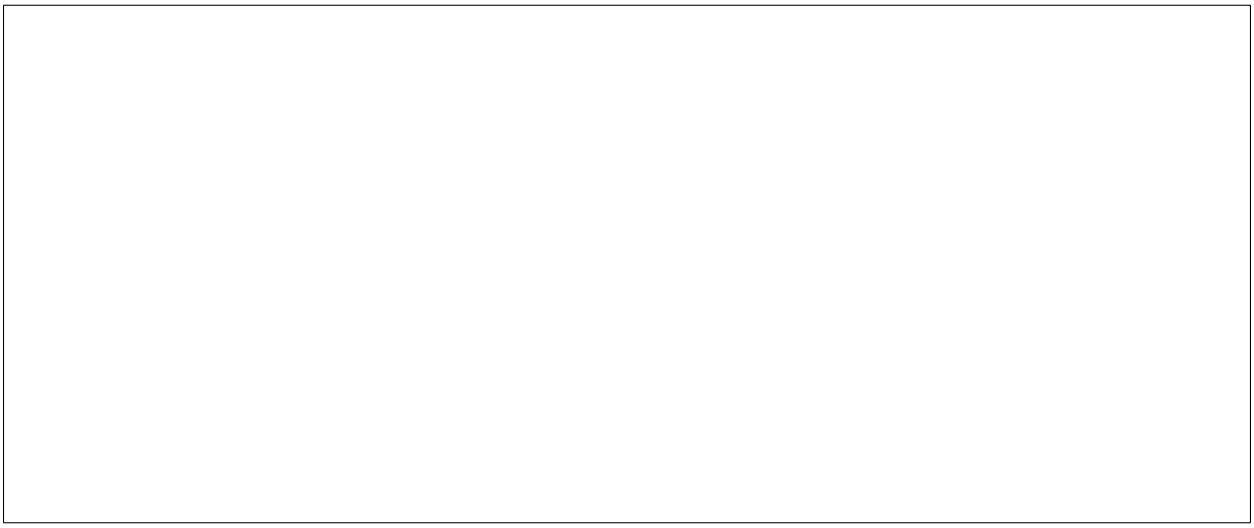
$$P(X_i|\theta) = \sum_{j=1}^{k} w_j P(X_i|r_j, \theta)$$

.

Both the r_j and w_j values can be estimated from the data. Unlike the gamma distribution model, the rate mixture model can approximate rate variation across sites regardless of the shape of this variation. For instance, the mixture model would be expected to improve dramatically on the gamma model if rate variation was distinctly bimodal or multimodal (Fig. 2).

Like the distributional models, mixture models can be applied to any substitution model parameter, even to tree parameters like branch lengths. Meade and Pagel (2004) give an interesting example

Figure 2: If rate variation across sites is bimodal, then a rate mixture model can improve considerably on a gamma distribution.



involving substitution model parameters. Kolaczkowski and Thornton (2004) discussed a mixture model on branch lengths but failed to implement it correctly.

A common problem for all mixture models is the identifiability of the processes. If we considered a mixture of two processes, a slow-rate and fast-rate process for instance, it is clear that the probability of the data would be the same regardless of whether the slow-rate process was assigned to process category 1 or category 2. To get around the problem of having to explore all possibilities, it is standard practice to introduce some constraint that guarantees a unique labeling of the processes. In the case of rate variation, for instance, we can simply order the rates such that $r_j < r_{j+1}$ for all j; we then know that the slow-rate process is always in category 1 and the fast in category 2.

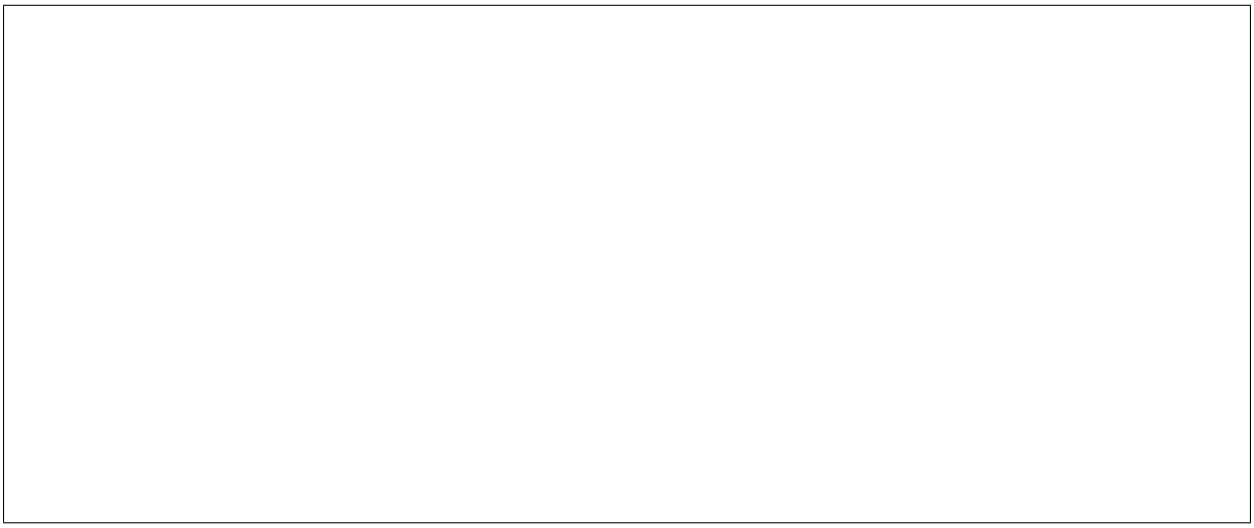
2.3 Hidden Markov Models

The approaches we have discussed thus far assume independence of sites but this is typically unrealistic. In particular, adjacent sites in a sequence tend to be much more likely to evolve under a similar process than two sites drawn at random. A popular way of accounting for the non-independence of adjacent sites is to model molecular evolution across a sequence using Hidden Markov Models (HMMs), perhaps best known for their application to speech recognition but now commonly used in many areas of science.

In a Hidden Markov Model, there is a discrete-state Markov process that moves the system be-

tween a number of *hidden states*. Each of these hidden states is associated with some *emission probabilities* with which observable data are generated when the system is in that state. For instance, assume that a molecular sequence consists of AT-rich regions with stationary state probabilities $\pi = \{0.4, 0.1, 0.1, 0.4\}$ and unbiased regions with stationary state probabilities $\pi = \{0.25, 0.25, 0.25, 0.25\}$; these would be the emission probabilities of the system in the AT-rich and the normal state, respectively. The crucial feature of the HMM is this distinction between the state of the system, which is hidden from direct observation, and the observable data, which are generated according to the emission probabilities of the hidden states. When we move from one site in the sequence to the next, the hidden states of the HMM change with the probability a_{ij} for the transition from state i to state j; assume that the probability of staying in an AT-rich region is 0.7 and the probability of staying in a normal region is 0.9. We can now define a HMM that allows us to identify AT-rich and normal sections of the sequence in terms of a *state transition diagram* (Fig. 3).

Figure 3: A simple HMM. It has two states, an AT-rich state emitting primarily adenine and thymine, and a normal state emitting all nucleotides with equal probabilities.



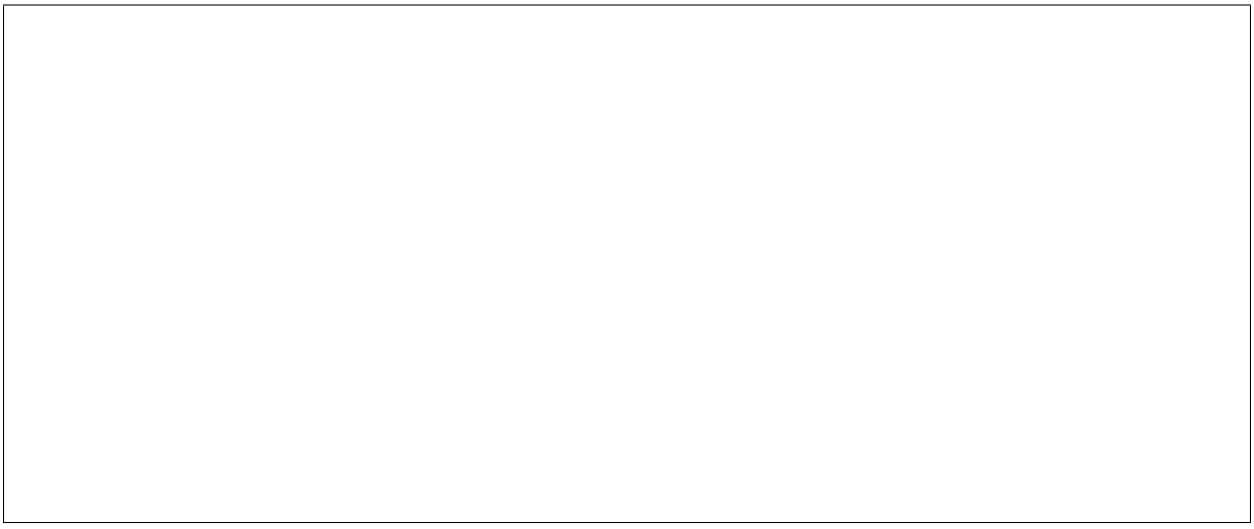
We start the HMM at the beginning of the sequence by drawing the initial state s_i from some probabilities b_i, typically the stationary probabilities of the different hidden states. In our case, these would be 0.25 for the AT-rich state and 0.75 for the normal state (since the rate of going in one direction is 3 times higher than the rate of going in the other direction). Assume that an observed sequence X of length n was generated by a path of hidden states ϕ, with the system being in state ϕ_i at sequence position i. The joint probability of this path and the observed sequence is

$$P(\phi, X) = b_{\phi_1} P(X_1|s_{\phi_1}) \prod_{i=2}^{n} a_{\phi_{i-1}\phi_i} P(X_i|s_{\phi_i})$$

We can calculate the total probability of the observed sequence by explicitly summing over all possible paths through the HMM. Since the number of paths grows exponentially with the number of sites in the sequence, this is not computationally feasible for real sequences. Fortunately, there is a dynamic programming algorithm known as the forward algorithm (not the backward algorithm as stated by Felsenstein (2004: 263); although the backward algorithm is essentially identical to the forward algorithm, it is used only in combination with the forward algorithm in typical HMM terminology), which allows us to compute the probability in time $O(nk^2)$, where n is the sequence length and k is the number of hidden states in the HMM. The algorithm is essentially identical to Felsenstein's pruning algorithm, which we will cover in the next lecture. We might also be interested in the most likely path through the HMM, which can be calculated using the so called Viterbi algorithm (analogous to the Sankoff algorithm), or the most probable state at each site, which is determined by the forward-backward algorithm (similar to the downpass-uppass algorithm combination of likelihood calculations on a tree, see next lecture).

We can easily apply HMMs to phylogenies by using the hidden states to model different parameter values of the substitution process. For instance, we can let the k states have different overall substitution rates, different GC content or vary in some other respect. Let us look at a simple variant of the rate HMM in which we have a single rate λ of moving between four hidden rate states (Fig. 4).

Figure 4: A phylogenetic HMM for rate variation across sites. It has k rate categories and a single rate λ of shifting between two states, an AT-rich state emitting primarily adenine and thymine, and a normal state emitting all nucleotides with equal probabilities.



This model can be made more complex by allowing the stationary state frequencies of the rate categories to be unequal, in analogy with the Markov models we have considered previously for

modeling the evolution of discrete data.

Since an HMM is based on a discrete-state discrete-time Markov process, the waiting times will be geometrically distributed (the discrete analog of the exponential distribution). This means that regions with a particular rate may have any length from very short to very long. In some type of HMMs, for instance HMMs of protein secondary structure, this distribution is not realistic. For instance, an alpha helix or beta sheet must involve at least a few amino acids, by definition. A work-around in such cases is to introduce a series of HMM states corresponding to an alpha helix or beta sheet and then force the HMM to go through all states before returning to another state. This type of solution is implemented in the PASSML model for protein secondary structure developed by Goldman and colleagues.

3 Heterogeneity across the tree

A standard assumption in phylogenetic models is process homogeneity across the tree. Again, models can be improved considerably by relaxing this assumption. This field is less developed than the modeling of process heterogeneity across sites but several approaches have been tried including branch breaking, heterogeneous substitution models, and compound Poisson process models.

3.1 Branch breaking

In the simplest case, we just break the phylogeny into two or more pieces and estimate substitution model parameters separately for each piece. This approach has been used extensively for detecting the presence of process heterogeneity across the tree. It requires that we have some idea *a priori* on the places where the tree should be cut.

An alternative approach is to let each branch has its own unique parameter value. However, this produces a large number of parameters and it does not capture the fact that, in most cases, we expect adjacent branches in the tree to evolve under the same process or at least very similar processes. Nevertheless, this approach can be very successful: the standard non-clock trees used in likelihood and Bayesian phylogenetic inference are good examples. In these trees, the overall rate is allowed to be unique for each branch, a good example of a model allowing process heterogeneity across the tree by treating branches separately.

A more sophisticated method is to assume some correlation in process parameter values between

adjacent branches. This is exemplified by Thorne and Kishino's method for accommodating rate variation across the tree, which will be covered in the lecture on tree models.

3.2 Heterogeneous substitution model

If we assume that process heterogeneity is independent for each site in the sequence, we can use substitution matrices that include hidden states. This type of approach is exemplified by the covarion and covariotide models. In the covariotide model, a site can be either in an on position, in which it evolves according to some standard four-by-four rate matrix $Q = \{q_{ij}\}$, or in the off position, in which it does not change at all. Assume that the rate of switching from the off position to the on position is s_{01} and the rate of switching in the other direction is s_{10}. The complete instantaneous rate matrix will then be an eight-by-eight matrix

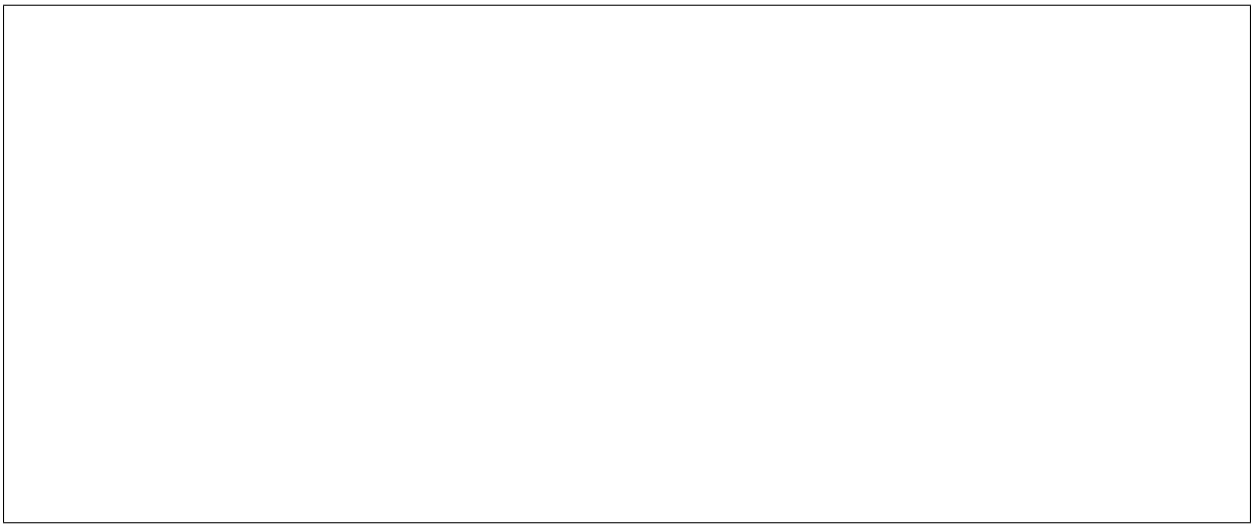
$$Q = \begin{array}{c} \\ \mathrm{A_{OFF}} \\ \mathrm{C_{OFF}} \\ \mathrm{G_{OFF}} \\ \mathrm{T_{OFF}} \\ \mathrm{A_{ON}} \\ \mathrm{C_{ON}} \\ \mathrm{G_{ON}} \\ \mathrm{T_{ON}} \end{array} \begin{array}{c} \begin{array}{cccccccc} \mathrm{A_{OFF}} & \mathrm{C_{OFF}} & \mathrm{G_{OFF}} & \mathrm{T_{OFF}} & \mathrm{A_{ON}} & \mathrm{C_{ON}} & \mathrm{G_{ON}} & \mathrm{T_{ON}} \end{array} \\ \left\{ \begin{array}{cccccccc} - & 0 & 0 & 0 & s_{01} & 0 & 0 & 0 \\ 0 & - & 0 & 0 & 0 & s_{01} & 0 & 0 \\ 0 & 0 & - & 0 & 0 & 0 & s_{01} & 0 \\ 0 & 0 & 0 & - & 0 & 0 & 0 & s_{01} \\ s_{10} & 0 & 0 & 0 & - & kq_{AC} & kq_{AG} & kq_{AT} \\ 0 & s_{10} & 0 & 0 & kq_{AC} & - & kq_{CG} & kq_{CT} \\ 0 & 0 & s_{10} & 0 & kq_{AG} & kq_{CG} & - & kq_{GT} \\ 0 & 0 & 0 & s_{10} & kq_{AT} & kq_{CT} & kq_{TG} & - \end{array} \right\} \end{array}$$

where k is a rate scaling constant determined by the proportion of time the sites spend in the on state. The covarion model is analogous, except that it is defined on the twenty amino acids and produces a forty-by-forty instantaneous rate matrix with twenty on states and twenty off states.

3.3 Compound Poisson process models

The compound Poisson process models are perhaps the most sophisticated of the models that have been tried for accommodating process heterogeneity across the tree. In a compound Poisson process, there is a basic Poisson process generating change points at which some parameter value changes according to a statistical distribution. For instance, we can let the overall rate of the substitution process change at these points by drawing a rate multiplier from an appropriate distribution (Huelsenbeck et al., 2000) at the change points. This particular compound Poisson process model will be described in more detail in the lecture on tree models.

Figure 5: Example of a compound Poisson process model of molecular evolution. Change points are generated by a Poisson process; at the change points, the value of some evolutionary process parameter is changed based on a random variable drawn from some appropriate distribution.



In principle, the compound Poisson process approach can be applied to any substitution model parameter although such attempts have not been published yet. An attractive feature of the model is that contiguous regions of the tree will evolve under the same model of evolution between change points, a model that many workers find plausible (Fig. 5).

4 Literature

A good summary of phylogenetic HMMs is provided by Siepel and Haussler (2005).