A Hidden Markov Model Approach to Variation Among Sites in Rate of Evolution

by

Joseph Felsenstein
and
Gary A. Churchill

BU-1304-M August 1995

A Hidden Markov Model Approach to Variation Among Sites in Rate of Evolution

Joseph Felsenstein*,[1] and Gary A. Churchill†,[2]

*Department of Genetics, University of Washington, Seattle, Washington 98195 and †Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York 14853

[1]Internet address: joe@genetics.washington.edu

[2]Internet address: gary@amanita.biom.cornell.edu

Running Head: Likelihood and evolutionary rates

Key words: likelihood, phylogeny, statistical inference, hidden Markov model, rate of evolution, molecular sequences

Corresponding Author:

Joe Felsenstein

Department of Genetics SK-50

University of Washington

Seattle, WA 98195, USA

Phone: (206) 543-0150

Fax: (206) 543-0754

Internet: joe@genetics.washington.edu

Abstract

The method of hidden Markov models is used to allow for unequal and unknown evolutionary rates at different sites in molecular sequences. Rates of evolution at different sites are assumed to be drawn from a set of possible rates, with a finite number of possibilities. The overall likelihood of a phylogeny is calculated as a sum of terms, each term being the probability of the data given a particular assignment of rates to sites, times the prior probability of that particular combination of rates. The probabilities of different rate combinations are specified by a stationary Markov chain that assigns rate categories to sites. While there will be a very large number of possible ways of assigning rates to sites, a simple recursive algorithm allows the contributions to the likelihood from all possible combinations of rates to be summed, in a time proportional to the number of different rates at a single site. Thus with 3 rates, the effort involved is no greater than 3 times that for a single rate. This "hidden Markov model" method allows for rates to differ between sites, and for correlations between the rates of neighboring sites. By summing over all possibilities it does not require us to know the rates at individual sites. However it does not allow for correlation of rates at non-adjacent sites, nor does it allow for a continuous distribution of rates over sites. It is shown how to use the Newton-Raphson method to estimate branch lengths of a phylogeny, and to infer from a phylogeny what assignment of rates to sites has the largest posterior probability. An example is given using β-hemoglobin DNA sequences in 8 mammal species; the regions of high and low evolutionary rates are inferred and also the average length of patches of similar rates.

Introduction

It has long been recognized that the assumption of equal rate of evolution implicit in many methods of analyzing phylogenies from molecular data is unrealistic. Maximum likelihood methods of inferring phylogenies from molecular sequences have always made this assumption (Neyman, 1971; Felsenstein, 1981). It is also implicit in almost all distance matrix methods using molecular sequences (e.g. Jukes and Cantor, 1969; Kimura, 1980). By assuming a given prior distribution of rates among sites one can correct these distance matrix methods for rate variation among sites (Olsen, 1987; Jin and Nei, 1990). However, such corrections do not restrict the effect of rate variation so that the same sites are inferred to have high rates of evolution across all members of the set of sequences. They also do not allow for any correlation in rates of evolution along the molecule.

Maxmimum likelihood methods can allow for variation in rates of evolution. For example, the PHYLIP package distributed by one of us (J.F.) has, in its DNAML and DNAMLK programs, versions 3.1 to 3.4, a "Categories" option that allows us to decide which rate category each site falls into, with the relative rates of evolution in different categories specified by us. This assumes that we know the relative rates of evolution in different sites, which is often not the case. Distance matrix methods can also be modified to allow for such site-specific rates, as is done in the program DNADIST in PHYLIP 3.5.

A halfways realistic treatment of rate variation among sites would have the following properties:

1. It must allow rates to differ among sites.

2. It must not assume that we know the relative rates of change at the individual sites, but must instead infer these from the data.

3. It must allow there to be some correlation between the rates of evolution at adjacent sites.

We will describe here a method of carrying out maximum likelihood estimation of phylogenies which satisfies these criteria. It will assume that there are a discrete set of possible rates (for example, one could assume that there were four different possible rates of evolution that stood in the ratios $0 : 1 : 2.3 : 8.9$). It will also assume that we can assign prior probabilities to these different rates, so that we feel able to say that the probability that a given site is in these four categories is (say) $0.10 : 0.32 : 0.22 : 0.36$. But we will not assume that we know which category of rate any given site is in. Furthermore we will allow correlation of rates among sites which are adjacent in the molecule.

We note that Yang (1993) and Kelly and Rice (1995) have developed methods of analyzing rate variation in a maximum likelihood analysis of phylogeny that satisfy conditions (1) and (2) above, and allow for a potentially infinite number of rate categories, so that we do not need to place any prior restriction on which rates are possible. This great generality is achieved at a cost: condition (3) is not met, and their calculations become difficult beyond a small number of species. Our approach will be less general in the rates it allows, but more general in allowing autocorrelation and in being useable in cases with many species. Yang (1994) has tested a similar discrete approximation, replacing a gamma distribution of rates by a discrete distribution with four well-chosen classes, and found it to perform well.

Our method uses the method of Hidden Markov Models (Baum and Petrie, 1966) which has been widely used in signal processing in communications, and was first applied to molecular sequences by Churchill (1989). Hidden Markov Models have also recently been applied to inferences of sequence alignment of proteins (Haussler et. al., 1993; Baldi et. al., 1994; Krogh et. al., 1994). Krogh et. al. also refer to

some other recent applications of Hidden Markov Models to molecular biology.

The method we describe requires an amount of computation that is greater than that for simple maximum likelihood inference of phylogenies by a factor roughly equal to the number of different rate categories. Thus, in the four-category example mentioned above, the amount of computation required to infer phylogenies is roughly four times as great as with a single rate. The method is implemented in versions 3.5 and later of the programs DNAML and DNAMLK in the PHYLIP package, which have been in distribution since March of 1993.

The present methods are quite similar to the auto-discrete-gamma model of Yang (1995), which was developed independently of them. He has used a bivariate Gamma distribution to model autocorrelation of rates among sites, and in order to effectively approximate this model has derived from it an autocorrelated Hidden Markov Model of rate variation. His model differs in detail from the present model but is similar in logical structure, and may give similar estimates of the phylogeny.

In this paper we outline the theory and computational methods for computing likelihood for a phylogeny with evolutionary rates that follow a Hidden Markov Model. We then explain the model of base substitution that is used in our implementation of this method, and the method of searching for the tree of highest likelihood, using a Newton-Raphson method that is specific to that base substitution model. We also also give a data example using mammalian hemoglobin sequences.

The model

The variation in evolutionary rates in our model is laid down by a Markov process that operates along the molecule, and assigns rates to sites. The rates are chosen from a finite pool of available rates, and the Markov process is assumed to be

stationary and irreducible, so that we can talk of the equilibrium probabilities f_i of the rates. The transition probabilities P_{ij} of this Markov process are assumed to be known. This Markov process is hidden from our view, as we cannot directly observe which sites evolve at which rates.

Once the sites have their rates assigned, each site will be assumed to evolve independently along the true phylogeny with that rate. Figure 1 depicts the model. Thus all correlation between sites will be assumed to be the consequence of the clustering (if any) of high and low rates at adjacent sites. A more complex model would be needed to deal with causes of correlation such as compensating substitutions in RNAs, both because the members of the pair of sites undergoing compensating substitutions may be widely separated along the molecule, and because the actual evolutionary events at the sites show a dependence that goes beyond their assignment to the same rate category.

The likelihood of a given phylogeny T is the sum, over all assignments of rate categories, of the probability of the data D given that combination of rates, multiplied by the prior probability of that combination of rates. If c_i denotes the category that a given rate combination assigns to site i, so that the rate assigned to site i is r_{c_i}, then if there are n sites we may write the likelihood of a given phylogeny as

$$L = \text{ Prob } (D|T) = \sum_{c_1} \sum_{c_2} \ldots \sum_{c_n} \text{ Prob } (c_1, c_2, \ldots c_n) \text{ Prob } (D|T, r_{c_1}, r_{c_2}, \ldots, r_{c_n}). \quad (1)$$

The assumption that each site evolves independently once the rate categories c_i are determined allows us to express the last probability as a product of terms, so that if D_i are the data at site i,

$$L = \sum_{c_1} \sum_{c_2} \ldots \sum_{c_n} \text{ Prob } (c_1, c_2, \ldots, c_n) \prod_{i=1}^{n} \text{ Prob } (D_i|T, r_{c_i}). \quad (2)$$

Simplifying the calculation

The hidden Markov model specifies that each combination of rate categories $c_1, c_2, ...c_n$ is the outcome of a stationary Markov chain, and thus its prior probability is simply the product of the prior probability of c_1 times a product of transition probabilities of that Markov chain:

$$\text{Prob}\,(c_1, c_2, ..., c_n) = f_{c_1} P_{c_1,c_2} P_{c_2,c_3} ... P_{c_{n-1},c_n} \tag{3}$$

It might be thought that there would be severe problems in computing (2), as, if there are k rate categories, the number of combinations of categories will be k^n. Thus with 1000 sites and 3 rate categories there are $3^{1000} \simeq 10^{477}$ terms to sum. In fact, the calculation can be done far more easily, using an algorithm that is similar in structure to the algorithm that calculates likelihood along a phylogeny. Let us denote by $D^{(k)}$ the data set consisting of sites k through n only. Then we can use (3) to write the likelihood as

$$L = \sum_{c_1} f_{c_1} \left(\sum_{c_2} \sum_{c_3} ... \sum_{c_n} \text{Prob}\,(c_2, ...c_n|c_1)\ \text{Prob}\,(D|T, r_{c_1}, r_{c_2}, ..., r_{c_n}) \right) \tag{4}$$

and then use (2) to rewrite it as

$$L = \sum_{c_1} f_{c_1} \left(\sum_{c_2} \sum_{c_3} ... \sum_{c_n} \text{Prob}\,(c_2, ...c_n|c_1) \prod_{i=1}^{n} \text{Prob}\,(D_i|T, r_{c_i}) \right). \tag{5}$$

The term in parentheses on the right-hand-side of (5) is the likelihood of the tree for $D^{(1)}$, given that site 1 has rate category c_1. Let us call this conditional likelihood $L^{(1)}_{c_1}$. We will, more generally, define $L^{(k)}_{c_k}$ as the likelihood of T for the data $D^{(k)}$ given that site k has rate category c_k. Then

$$L = \sum_{c_1} f_{c_1} L^{(1)}_{c_1}. \tag{6}$$

We can use equation (2) to write

$$L_{c_1}^{(1)} = \text{Prob}\,(D_1|T, r_{c_1}) \sum_{c_2} \sum_{c_3} \ldots \sum_{c_n} \text{Prob}\,(c_2, \ldots c_n|c_1) \prod_{i=2}^{n} \text{Prob}\,(D_i|T, r_{c_i}). \quad (7)$$

Equation (3) now allows us to write the conditional probability of $c_2, c_3, \ldots, c_n$ given c_1 as $P_{c_1,c_2} P_{c_2,c_3} \ldots P_{c_{n-1},c_n}$ which allows us to rewrite $L_{c_1}^{(1)}$ as

$$L_{c_1}^{(1)} = \text{Prob}\,(D_1|T, r_{c_1}) \sum_{c_2} P_{c_1,c_2} \left(\sum_{c_3} \ldots \sum_{c_n} \text{Prob}\,(c_2, \ldots c_n|c_1) \prod_{i=2}^{n} \text{Prob}\,(D_i|T, r_{c_i}) \right). \quad (8)$$

Noting that the expression in parentheses on the right-hand-side of (8) is just $L_{c_2}^{(2)}$, we have an expression for the $L_{c_1}^{(1)}$ in terms of the $L_{c_2}^{(2)}$:

$$L_{c_1}^{(1)} = \text{Prob}\,(D_1|T, r_{c_1}) \sum_{c_2} P_{c_1,c_2} L_{c_2}^{(2)}. \quad (9)$$

This suggests that a general recursion might exist, calculating each of the $L_{c_k}^{(k)}$ in terms of the $L_c^{(k+1)}$, and in fact this is easily shown by continuing the same argument, repeatedly using (2) and (3), that

$$L_{c_k}^{(k)} = \text{Prob}\,(D_k|T, r_{c_k}) \sum_{c_{k+1}} P_{c_k,c_{k+1}} L_{c_{k+1}}^{(k+1)}. \quad (10)$$

The exception to this equation is when $k = n$, in which case by definition

$$L_{c_n}^{(n)} = \text{Prob}\,(D_n|T, r_{c_n}). \quad (11)$$

The pattern of computation reverses the order of the recursion in equation (10). First, we must compute all the Prob $(D_k|T, r_{c_k})$, which are the likelihoods at each site for each possible rate category. The amount of computation for this will be proportional to the product of the number of sites and the number of rate categories. Then we use (11) to determine the values of the $L_{c_n}^{(n)}$. Then (10) can be used to compute $L_{c_{n-1}}^{(n-1)}, L_{c_{n-2}}^{(n-2)}$, and so on down to $L_{c_1}^{(1)}$. There are $n-1$ steps in this computation, each one in the most general case requiring an effort proportional to

the square of the number of rate categories. Finally equation (6) is used to compute L. The storage requirements of this computation are modest: we can store all of the values of Prob $(D_k|T, r_{c_k})$, there being n times the number of rate categories of these. The computation can be done with less storage than this, although in most cases that economy will be unnecessary.

The computation described here proceeds from the last site, n to the first one. It could just as easily be done in the other direction, in which case the formulas would be analogous, P_{ij} being replaced by the reverse transition probability Q_{ji}, where we have the usual formula for computing the transition probabilities for the reversed Markov chain

$$Q_{ji} = f_i P_{ij} / f_j. \tag{12}$$

If the Markov chain is reversible, then the Q_{ij} and the P_{ij} will be identical.

As stated here, the computation may require effort proportional to the square of the number of rate categories. However, for the particular choice of P_{ij} used in our implementation of this method, described below, the computation in equation (10) can be done in a time linear in the number of rate categories.

The most probable combination of rates

Our ability to calculate the likelihood of the phylogeny T allows us to search for the maximum likelihood phylogeny. Once that is estimated, we may want to see some indication of what the rates of evolution are at the different sites. The likelihood has been computed by summing contributions from all possible combinations of rates. One combination that may be of particular interest is the combination that makes the largest contribution to the likelihood. This will depend on both the prior probability of the combination and the likelihoods at the sites, as its contribution

will be:

$$R = \max_{c_1,c_2,...,c_n} \text{Prob}\,(c_1, c_2, ...c_n)\ \text{Prob}\,(D|T, r_{c_1}, r_{c_2}, ..., r_{c_n}). \tag{13}$$

There is an algorithm, closely related to the one used to sum likelihoods in the previous section, that finds the combination of rates that achieves this maximum. It is a version of the algorithm of Viterbi (1967), which is well explained by Forney (1973). In an analogue to the quantity $L_{c_i}^{(i)}$ of the previous section, let us define $R_{c_k}^{(k)}$ as the likelihood contribution for sites k through n for the combination of rates that has site k having rate category c_k, and sites $k+1$ through n having that combination of categories that maximizes the contribution of sites k through n, so that we define

$$R_{c_k}^{(k)} = \max_{c_{k+1},...,c_n} \text{Prob}\,(c_{k+1}, ...c_n|c_k)\ \text{Prob}\,(D^{(k)}|T, r_{c_k}, r_{c_{k+1}}, ..., r_{c_n}). \tag{14}$$

For $k = n$ the definition (14) specifies that

$$R_{c_n}^{(n)} = \text{Prob}\,(D_n|T, r_{c_n}) \tag{15}$$

which we have already calculated. For all other values of k we have a relation analogous to (10), but taking maxima rather than summing the contributions:

$$R_{c_k}^{(k)} = \text{Prob}\,(D_k|T, c_k) \max_{c_{k+1}} \left[P_{c_k,c_{k+1}} R_{c_{k+1}}^{(k+1)}\right]. \tag{16}$$

Using this successively on sites $n-1$, $n-2$, and so on down to 1, we end up with the $R_{c_1}^{(1)}$ for all possible categories c_1 for site 1. The largest of the quantities $f_{c_1} R_{c_1}^{(1)}$ is the size of the largest contribution of a single combination of rate categories to the likelihood.

This leaves us without yet knowing the combination of categories $c_1, c_2, ..., c_n$ that achieved this maximum. However, as we used equation (16) for each site we computed, for each possible rate category at that site, the rate category c_{k+1} at the next site that maximized the contribution. Suppose that we call this $C_{c_k}^{(k)}$, so that

$C_{c_k}^{(k)}$ is the value of c_{k+1} that is selected by the maximization in equation (16). These values of c_{k+1} can be stored in the array S as the computation proceeds from site n down to site 1. At the end we know which rate category c_1 corresponds to the maximum contribution. We can then use $C_{c_1}^{(1)}$ to find the value of c_2 that is involved in the maximum contribution, and then $C_{c_2}^{(2)}$ specifies the category for site 3, and so on. Backtracking in this way we quickly read off the combination of rate categories that makes the largest contribution, and report these.

The combination of rate categories that makes the largest contribution to the likelihood is not necessarily the only one that might be of interest. We might also imagine finding, for each site, the rate category at that site that is involved in making the largest total contribution to the likelihood (so that the sum of the contributions of all combinations of rate categories that have category c_k at site k is as large as possible). If for each combination of rate categories we divide their contribution to the likelihood by the overall likelihood, these quantities will sum to 1, and we can consider them as a probability distribution. The quantity R we were computing in equations (14)-(16) is the size of the mode of that distribution. The present quantity is in effect for each site k the mode of the marginal distribution over c_k. In general, the categories that together make the largest contribution to the likelihood will usually also be the ones that individually make the largest site-by-site marginal contribution, but there can be cases in which the two methods will select different combinations of rates. We will see below that it is not hard to compute the combination of rate categories that have the largest marginal contributions, using an algorithm similar to those given above, but making two passes through the sites, one from n down to 1, and one back up again.

The implementation

The discussion above applies to any stationary Markov process for assigning rates to sites, and any Markov process that has such rates as a parameter and that controls the evolution of sites at those rates on a given phylogeny. The hidden Markov model method for allowing for rate variation has been implemented in version 3.5 of the programs DNAML and DNAMLK in the PHYLIP Phylogeny Inference Package, which is distributed by one of us (J.F.) and is available for free, including distribution over Internet by anonymous ftp from evolution.genetics.washington.edu. This version was first made available in March, 1993. While we emphasize that the general method applies to many other models, in this section we will give some details of the particular models used in these programs.

We are allowed to specify the number of different rate categories that will be possible, the relative rates r_i of the different categories, and the equilibrium probabilities f_i of each category. The r_i may be any nonnegative real numbers, and the f_i any set of frequencies that add to 1. Note that we can allow for invariant sites by simply having one category that has $r_i = 0$. There also is an autocorrelation parameter, which we will call λ. Each site is assumed to have a probability λ that the rate at that site is the same as at the previous site. With probability $1 - \lambda$ the rate is instead drawn at random from the equilibrium distribution of rates, including the possibility that the same rate is drawn again by chance.

It is possible to estimate the values of the relative rates r_i and probabilities f_i and the autocorrelation parameter λ, using the EM algorithm of Baum et al. (1970). However implementation of this algorithm for the rates would significantly increase computation required (it could more readily be used to estimate the autocorrelation parameter alone). We have found in practice that it is more efficient to examine a few sets of rates and correlation values and choose one that yields the highest

likelihood.

The transition probability, under this model, from state i to state j will be

$$P_{ij} = \lambda\delta_{ij} + (1-\lambda)f_j \tag{17}$$

where δ_{ij} is the Kronecker delta function, which is 1 when $i = j$ and 0 otherwise. This model will achieve the stated equilibrium distribution f of rate categories. If λ is near 1 there will be a large autocorrelation of rate categories among neighboring sites; if it is 0 there will be no autocorrelation. The expected size of a patch of sites would be $1/(1-\lambda)$, except that there is nothing in this model that prevents the next rate category that is chosen from being the same as the present one. The overall probability that the rate does not change from one site to the next is the weighted average of the P_{ii}:

$$\sum_i f_i P_{ii} = \lambda + (1-\lambda)\sum_i f_i^2. \tag{18}$$

and this value can be used to compute the mean apparent patch size. If there are two rate categories of equal frequency, this number is $(\lambda + 1)/2$. If there are 10 categories of equal frequency, it is $(0.9\ \lambda + 0.1)$, which is much closer to the value of λ that would hold if adjacent patches never accidentally had the same rate. In the DNAML and DNAMLK implementations, we are asked to specify an "average patch length", but this is actually taken to be $1/(1-\lambda)$, and λ is set from its value. In view of equation (18) this will be slightly incorrect.

The model of base change used in the programs

The computational scheme presented above will work for any model of base change for which we can specify evolutionary rates that may differ from site to site. In most models this is easily done by allowing the branch lengths in the phylogeny to be

proportional to the rates of evolution (and thus to differ from site to site). In effect, we treat a site that has twice the rate of evolution as if it evolves along a branch that is twice as long. Thus if we have a model of evolution that has a transition probability that depends on both branch length t and evolutionary rate r so that it is $M_{ij}(t,r)$, the rates can be accomodated by multiplying the branch length by r if and only if

$$M_{ij}(t,r) = M_{ij}(rt,1) \tag{19}$$

This is true for most models of base change, as they accomodate site-specific rates of evolution by replacing the time t by the product $r_k t$ for site k.

The particular model that we have used in DNAML and DNAMLK version 3.5 is one that allows for inequalities of equilibrium base composition and for inequalities of the rate of transitions and transversions. It is related to the model given by Felsenstein (1981) but generalizes it to allow for unequal rates of transitions and transversions. Hasegawa, Kishino and Yano (1988; also Kishino and Hasegawa, 1989) have previously described this model in print, in the course of describing their own model that also allows for inequalities of base composition and transition/transversion rates. Their model is similar to the present one but not identical to it; in practice the similarity was such that they were willing to use the present model in many of their likelihood computations using programs from the PHYLIP package. A similar but not identical model has also been developed by Rempe (1988). The present model has been used by J.F. in versions of the PHYLIP package distributed since 1984.

The model can be described as having two kinds of event, I and II. The first can generate either no change or a transition, the second no change, a transition, or a transversion. Suppose that the rates of these two events are called α and β. Event I is the replacement of the nucleotide at the site by one that is randomly

sampled from the equilibrium pool of purines (if the original base is a purine) or pyrimidines (if the original base is a pyrimidine). For example, a base which is an A is replaced by another A with probability $\pi_A/(\pi_A + \pi_G)$, and with a G with probability $\pi_G/(\pi_A + \pi_G)$. A base which is a C is replaced by another C with probability $\pi_C/(\pi_C + \pi_T)$, and with a T with probability $\pi_T/(\pi_C + \pi_T)$. Thus an event of type I may either cause no change or a transition.

An event of type II replaces the base with one drawn from the pool of all four possible nucleotides, with probabilities equal to their equilibrium base composition. Thus an A is replaced by another A with probability π_A, by a G with probability π_G, by a C with probability π_C, and by a T with probability π_T. An event of type II can case no change, a transition, or a transversion.

The overall rate of substitution per site will be

$$\begin{aligned}\mu = \alpha\left(\pi_A\left(\tfrac{\pi_G}{\pi_A+\pi_G}\right) + \pi_G\left(\tfrac{\pi_A}{\pi_A+\pi_G}\right) + \pi_C\left(\tfrac{\pi_T}{\pi_C+\pi_T}\right) + \pi_T\left(\tfrac{\pi_C}{\pi_C+\pi_T}\right)\right) \\ +\beta(\pi_A(1-\pi_A) + \pi_G(1-\pi_G) + \pi_C(1-\pi_C) + \pi_T(1-\pi_T)).\end{aligned} \tag{20}$$

If π_R and π_Y are the equilibrium base frequencies of purines and pyrimidines, respectively, so that

$$\pi_R = \pi_A + \pi_G \tag{21}$$

and

$$\pi_Y = \pi_C + \pi_T, \tag{22}$$

then we can simplify (20) to become

$$\mu = \alpha\left(2\pi_A\pi_G/\pi_R + 2\pi_C\pi_T/\pi_Y\right) + \beta\left(1 - \pi_A^2 - \pi_G^2 - \pi_C^2 - \pi_T^2\right). \tag{23}$$

The ratio of transitions to transversions will be

$$R = \left(\alpha\left(2\pi_A\pi_G/\pi_R + 2\pi_C\pi_T/\pi_Y\right) + \beta\left(2\pi_A\pi_G + 2\pi_C\pi_T\right)\right) / \left(\beta 2\pi_R\pi_Y\right). \tag{24}$$

Expressing the instantaneous rates of transition b_{ij} between the different nucleotides in terms of the rates α and β of type I and type II events we get for any two bases i and j

$$b_{ij} = -\delta_{ij}(\alpha + \beta) + \epsilon_{ij}\alpha\frac{\pi_j}{\sum_k \pi_j \epsilon_{jk}} + \beta\pi_j, \tag{25}$$

where δ_{ij} is the usual Kronecker delta function, and ϵ_{ij} is a similar function which is 1 when i and j are either both purines or both pyrimidines, and 0 otherwise. Note that the term $\sum_k \pi_j \epsilon_{jk}$ simply computes either π_R or π_Y, depending on whether j is a purine or a pyrimidine. This parameterization of the model is essentially the same as that given by Hasegawa, Kishino, and Yano (1988).

Solving (23) and (24) for α and β, we get

$$\alpha = \frac{2\pi_R\pi_Y R - (2\pi_A\pi_G + 2\pi_C\pi_T)}{(2\pi_A\pi_G/\pi_R + 2\pi_C\pi_T/\pi_Y)}\frac{\mu}{1+R} \tag{26}$$

and

$$\beta = \frac{1}{2\pi_R\pi_Y}\frac{\mu}{1+R}. \tag{27}$$

We can express the instantaneous rates (25) in terms of μ and R by substituting (26) and (27) into (25). The results are straightforward and not particularly edifying and we will not give them here.

An advantage of the present model is that it is easy to compute transition probabilities for any time t. If there has been at least one event of type II during this time, the probability of resulting base being j is simply π_j, regardless of how many other events of either type have also occurred. If there has been no event of type II but at least one event of type I, the probability of the resulting base being j is simply $\pi_j/\sum_k \pi_j \epsilon_{ik}$, regardless of how many other events of type I have occurred. As the probability of at least one event of type II is $1 - \exp(-\beta t)$, and the probability of no event of type II but at least one of type I is $\exp(-\beta t)(1 - \exp(-\alpha t))$, the transition

probabilities can be given as

$$M_{ij}(t,1) = e^{-(\alpha+\beta)t}\delta_{ij} + e^{-\beta t}\left(1 - e^{-\alpha t}\right)\left(\frac{\pi_j}{\sum_k \pi_j \epsilon_{ik}}\right)\epsilon_{ij} + \left(1 - e^{-\beta t}\right)\pi_j, \qquad (28)$$

and they can be re-expressed in terms of the more meaningful parameters μ and R by substituting from (22) and (23) for α and β.

Evaluating the likelihoods along the tree

Given that we can evaluate the likelihood of any given tree T for any given parameter value λ, we still have to solve the problem of maximizing the likelihood over all T and all λ. In practice the methods used in DNAML version 3.5 and DNAMLK version 3.5 are not sophisticated. Many of the particulars have been described earlier (Felsenstein, 1981) although the program in current distribution differs in many ways from that described in 1981. For a given phylogeny in DNAML each branch length is iterated separately (in DNAMLK each ancestral node time is iterated separately), using the Newton-Raphson method, repeatedly evaluating the likelihood. This does not require a re-evaluation of likelihoods throughout the tree each time, because the "pruning" algorithm can be used.

This algorithm, a relative of the "peeling" algorithm in statistical human genetics, has been described by Felsenstein (1973, 1981), but a brief review here will be useful. Suppose that we define $\ell_{ic}^{(m)}(s)$ as the likelihood of the tree for all data for site m at or above node i on the tree, given that site m in node i is in state s, and given that site m has rate category c. We can easily determine this for the tips of the tree. If, for example, tip i shows an A in site m, it follows immediately by its definition that $\ell_{ic}^{(m)}(A) = 1$, and the ℓ value for all three other bases b is $\ell_{ic}^{(m)}(b) = 0$. We can work down the tree computing ℓ values at each site for each node of the tree, by making use of the recursion for a node i whose immediate descendants, j and k,

have ℓ values that have been previously computed, and have branch lengths v_j and v_k leading to them:

$$\ell_{ic}^{(m)}(s) = \left[\sum_{x=A}^{T} M_{sx}(v_j,1)\ell_{jc}^{(m)}(x)\right]\left[\sum_{y=A}^{T} M_{sy}(v_k,1)\ell_{kc}^{(m)}(y)\right]. \tag{29}$$

This process proceeds down the tree towards the root. In an unrooted tree the root may be taken to be anywhere. The values of $\ell_{ic}^{(m)}(s)$ at the root are then combined in a weighted average

$$L_c^{(m)} = \sum_{x=A}^{T} \pi_x \ell_{ic}^{(m)}(x) \tag{30}$$

which computes the likelihood at that site for the whole tree, for rate category c, unconditioned on knowing the base at that node.

Branch lengths by the Newton-Raphson method

The preceding process allows us to compute site- and rate-category-specific likelihoods for the nodes at both ends of any given branch, by simply assuming the root to be in that branch and "pruning" the likelihoods from the tips down until they arrive at the nodes at the two ends of the branch. We can then use these to find the length of that branch that optimizes the likelihood. In PHYLIP 3.5 we did this by a simple, and excessively slow, line search of the branch lengths, using (29), (30), (10) and (11) to compute the overall likelihood for each branch length. It was accelerated somewhat by making a quadratic prediction of the optimal branch length after every three steps of the line search.

In PHYLIP 3.6 this process is replaced by the Newton-Raphson method, which is considerably faster. We could have done simultaneous Newton-Raphson iteration of all branch lengths. This might have been better but was computationally tedious. We have instead opted to iterate the lengths of one branch at a time. Appendix A

shows the calculations of the first and second derivatives needed for this iteration. The equations for computing them can be obtained by taking derivatives in equations (10)-(11).

Appendix A presents the formulas generally. In Appendix B shows the calculations, for the particular model of DNA change used in DNAML, of the quantities Prob $(D_k|T, r_{c_k})$, d Prob $(D_k|T, r_{c_k})/dv$ and d^2 Prob $(D_k|T, r_{c_k})/dv^2$.

These derivatives are used in the formulas (6), (10), (11), (A1) - (A3), (A4)-(A6) to obtain the derivatives of the likelihoods in a recursive calculation along the sequence. In DNAML from PHYLIP 3.6 these derivatives are used to estimate the branch length by use of the Newton-Raphson method. This is modified so that it always moves in an uphill direction; if it overshoots, points 1/2, 1/4, 1/8... of the way are tried successively until one finally results in an increase in the likelihood.

Traversing through the tree, branch lengths are successively optimized until an adequate number of traversals has occurred. At that point the best branch lengths and likelihood are available for the given tree topology. The search among tree topologies is conducted, in the terms of Swofford and Olsen (1990) by stepwise addition followed by branch-swapping by nearest-neighbor interchanges after each species is added. A final round of branch swapping by subtree pruning and regrafting is available as an option. So are multiple runs with different input orders of species, the tree reported being the best one found among all the runs with different input orders.

It is also possible to estimate branch lengths by the EM algorithm (Dempster, Laird, and Rubin, 1977), but we will not go into details about that here.

Regional and site-specific rates

The preceding sections have explained how we can construct a method that allows rate variation from site to site in an autocorrelated pattern, in which it is not known in advance which sites will have high or low rates. However, this leaves us without a way to analyze data where there are codon site-specific rate variations. If we know which nucleotide sites are the first, second, and third positions in the codons, we would like to be able to specify that these vary in rate of evolution, while also allowing regional rate variation.

The simplest approach to this, used in version 3.6 of DNAML and DNAMLK, is to let the rate at each site be the product of two rates, one of which is the site-specific rate that we have specified, and the other of which is the rate assigned by the Hidden Markov Model. In these programs we are asked to specify a number of rate categories, their rates of evolution, and to assign each site to a category. Thus a (tiny) protein with a short intron might have site-specific categories 1, 2, 3, and 4, with 4 being the category for intron sites. We might preassign categories

123123123123123123123123123123444444

444444444444444444123123123123123123123

and also allow regional rate variation to be inferred by the Hidden Markov Model methods we have outlined above. The computations are no harder – we just make sure when computing the quantities Prob $(D_i|T, r_{c_i})$ to have the rate for site i with regional rate category c_i be not r_{c_i}, but $\rho_i r_{c_i}$, where ρ_i is the preassigned rate for site i. Thereafter the computations go through as we have outlined, without any additional computations.

A product of rates is, however, not entirely realistic. If third positions of codons are allowed to have a high rate of evolution because they are nearly unconstrained

by natural selection against mutants, they will not necessarily have a higher rate of evolution in parts of the molecule that are under less constraint. A more realistic assumption would be a saturation function such as

$$1 - e^{-\rho_i r_{c_i}} \tag{31}$$

or

$$\rho_i r_{c_i} \Big/ (1 + \rho_i r_{c_i}) . \tag{32}$$

With these functions, a third position might have a much higher rate of evolution than a second position if we are in a highly constrained region of the protein, but it might have only slightly greater rate of evolution in little-constrained region. We hope to implement such a saturation function in future versions of PHYLIP, if we can do so without confusing users.

Possible future extensions

The growing use of Hidden Markov Models in protein structure modelling suggests that it should be possible to combine those structural HMM's with the ones we use here. The main difficulty in doing so is that the hidden states in protein structure modelling are correlated not only along the molecule, but spatially as well. For example, in RNA secondary structures, sites that are well-separated in the linear sequence may be part of the same loop. To properly model the evolutionary rates of sites in the loop, we would need to allow the hidden states to be correlated spatially, not simply autocorrelated along the molecular sequence.

The present framework also does not allow the changes themselves to be correlated. Compensating substitutions are a major source of information about secondary structure in RNAs, and may be of comparable use in proteins. The

present models allow two sites to have correlated rates, but once those rates are assigned there is then assumed to be independent change at the two sites. Tillier (1994; Tillier and Collins, 1995) has modelled RNA base-pair substitution using a six-state model (AU, GU, GC, UA, UG, CG) with 7 parameters. This constrains the substitution events to be correlated. It would be of great interest to combine her approach with Hidden Markov Models of stem and loop states, particularly if a way can be found to represent the pairing of states in the HMM. Of course, the same problems, and opportunities, exist for proteins, although the difficulties are expected to be greater.

In addition to states representing structure, we might have states representing expected purine/pyrimidine content. One state might represent being in an AT-rich region, the other being in a GC-rich region. The mathematics involved is essentially identical to that outlined above, except that the transition probability matrix M_{xy} used in equations (29), (B1), and (A1) would differ between AT-rich and GC-rich states, by having a different equilibrium distribution of nucleotides. Multiple AT-rich and GC-rich states could be used to model different base composition states. How many different states will be needed to realistically model base composition variation is not known.

Hidden Markov models could be developed to detect change points in the tree topology along the length of a set of aligned sequences. Topology changes can result from recombination, gene conversion or horizontal transfer events that may have occurred within the history of the sequences. The methods developed by Hein (1993) are based on parsimony rather than likelihood methods but they make use of algorithms similar to the Hidden Markov Model algorithms. The states of the hidden Markov model in this case would be tree topologies and thus the number of states may be unmanageably large for even moderate numbers of sequences. The problems

of detecting and modeling recombination events will become increasingly important as more within species sequence samples are collected. A likelihood based approach to modeling recombination is described by von Haeseler and Churchill (1993).

A data example

To illustrate the technique, we have collected from Genbank release 82 the coding sequences (omitting introns and flanking sequences) of eight mammalian β-hemoglobins. Their species names and accession numbers are: *Tachyglossus aculeatus* (L23800), *Didelphis virginiana* (J03643), *Capra hircus* (M15387), *Rattus norvegicus* (M17084), *Oryctolagus cuniculus* (K03256), *Tarsius syrichta* (J04429), *Lemur macaco* (M15734), *Homo sapiens* (U01317). These have been aligned using ClustalV (Higgins and Sharp, 1989), which is easily done; only two gaps have to be introduced. A series of runs has been done with site-specific categories representing the three codon positions, and with two regional rates. The best combination of parameters that has been found so far has rates 1.0 : 0.6 : 2.7 for the codon-position relative rates, and rates 1.0 : 8.0 for the two regional rates. The frequencies of the two regions are inferred to be 0.75 : 0.25, and the parameter λ is inferred to be 0.5454, which means that one expects to choose a new regional rate every 2.2 bases on average.

The phylogeny is shown in figure 2. It is outgroup-rooted on the branch leading to the echidna *Tachyglossus*, and shows the opossum *Didelphis* branching off next, and the placental mammals as a monophyletic group. The positions of the rat *Rattus*, lemur, and rabbit *Oryctolagus* are of dubious correctness, but the branches defining this structure are small. When a likelihood ratio test is made of those branches by holding them to length zero while optimizing the lengths of all other branches, they

each prove to be statistically insignificant, in neither case resulting in a reduction of more than 0.5 units of log-likelihood. By the same method the group of placental mammals and the separation of the goat *Capra* from the other placental mammals both prove to be significant, leading to a drop of more than 8 units of log-likelihood when these branches are forced to have length zero. We should note that

Of greater interest will be the inferences about which regions have high and low rates of change. Figure 3 shows the sequences, using the dot-differencing convention according to which a dot means "the same as the first sequence". Below each block of 50 bases is shown the two inferences of rates. The upper line of 1's and 2's shows the combination of regional rates which makes the largest contribution to the likelihood. The line below it shows a 1 or a 2 when the fraction of all likelihood that is accounted for by rate combinations that have a 1 or 2 in that position is more than 95%. Otherwise it shows a space.

Certain features are unsurprising, such as conservation of the codon for the heme-binding Histidines (sites 190-192 and 277-279). The eight α-helical regions of the protein (A: sites 13-60, B: 61-105, C: 106-126, D: 151-171, E: 172-231, F: 259-282, G: 298-354, and H: 370-435) show within them patches of high and low rates. What is more striking is that in the non-helical regions (all the remaining ones except sites 1-3 whose amino acid product does not appear in the final protein), there are markedly fewer high rates than low. In the rate combination that is most probable a posteriori the helical regions have a high rate in 101 out of 244 sites, but the nonhelical regions only in 14 out of 82 sites. Though not easily statistically testable, this fits in with the notion that the helical regions are under less constraint than the nonhelical ones.

Acknowledgements

Research by J.F. was supported by grants number BSR-8918333 and DEB-9207558 from the National Science Foundation, and grants number 1 R01-GM41716-01 and 2 R55 GM41716-04 from the National Institute of General Medical Sciences, National Institutes of Health. Research by G.C. was supported by grant number DE-FG02-93ER61567 from the Department of Energy. We wish to thank Jeff Thorne, Ziheng Yang, Mary Kuhner, and Peter Beerli for discussion and comments on the manuscript.

Appendix A

Derivatives of likelihoods for the Newton-Raphson method

For the Newton-Raphson iteration of a branch length v one needs the first and second derivatives of the likelihood. The first can be computed from (10)-(11) by taking the derivatives with respect to branch length of the likelihoods on their left-hand-sides:

$$\frac{dL_{c_k}^{(k)}}{dv} = \frac{d \text{ Prob } (D_k|T, r_{c_k})}{dv} \sum_{c_{k+1}} P_{c_k,c_{k+1}} L_{c_{k+1}}^{(k+1)} + \text{ Prob } (D_k|T, r_{c_k}) \sum_{c_{k+1}} P_{c_k,c_{k+1}} \frac{dL_{c_{k+1}}^{(k+1)}}{dv} \tag{A1}$$

and

$$\frac{dL_{c_n}^{(n)}}{dv} = \frac{d \text{ Prob } (D_n|T, r_{c_n})}{dv}. \tag{A2}$$

Using (A2) and (A1) we can recursively compute the quantities $dL_{c_k}^{(k)}$ from $k = n$ down to $k = 1$. The derivative of the overall likelihood with respect to the branch length is simply, from (6)

$$\frac{dL}{dv} = \sum_{c_1} f_{c_1} \frac{dL_{c_1}^{(1)}}{dv}. \tag{A3}$$

Similarly, we can compute the second derivative of the likelihood with respect to the branch length by differentiating again, getting

$$\begin{aligned} \frac{d^2 L_{c_k}^{(k)}}{dv^2} &= \frac{d^2 \text{ Prob } (D_k|T,r_{c_k})}{dv^2} \textstyle\sum_{c_{k+1}} P_{c_k,c_{k+1}} L_{c_{k+1}}^{(k+1)} \\ &+ 2 \frac{d \text{ Prob } (D_k|T,r_{c_k})}{dv} \textstyle\sum_{c_{k+1}} P_{c_k,c_{k+1}} \frac{dL_{c_{k+1}}^{(k+1)}}{dv} \\ &+ \text{ Prob } (D_k|T, r_{c_k}) \textstyle\sum_{c_{k+1}} P_{c_k,c_{k+1}} \frac{d^2 L_{c_{k+1}}^{(k+1)}}{dv^2}, \end{aligned} \tag{A4}$$

and

$$\frac{d^2 L_{c_n}^{(n)}}{dv^2} = \frac{d^2 \text{ Prob } (D_n|T, r_{c_n})}{dv^2}, \tag{A5}$$

and at the end

$$\frac{d^2 L}{dv^2} = \sum_{c_1} f_{c_1} \frac{d^2 L_{c_1}^{(1)}}{dv^2}. \tag{A6}$$

Thus the quantities $L_{c_k}^{(k)}$, $dL_{c_k}^{(k)/dv}$, and $d^2 L_{c_k}^{(k)}/dv^2$ can be computed recursively by proceeding from $k = n$ down to $k = 1$, and at the end the values for $k = 1$ can be combined using (6), (A3), and (A6) to get the likelihood and its first and second derivatives with respect to this branch length.

Appendix B

Derivatives of sitewise likelihoods in DNAML

In the DNAML program the quantities Prob $(D_k|T, r_{c_k})$, d Prob $(D_k|T, r_{c_k})/dv$ and d^2 Prob $(D_k|T, r_{c_k})/dv^2$ are obtained by taking the root of the tree to be at the node (j) at one end of the branch, the node at the other end being node k. If the length of the branch is v_k, the overall likelihood at site i given that the rate category for that site is c_i is

$$\text{Prob } (D_i|T, r_{c_i}) = \sum_x \pi_x \ell^{(i)}_{jc_i}(x) \sum_y M_{xy}(v_k, r_{c_i}) \ell^{(i)}_{kc_i}(y), \tag{B1}$$

and the first and second derivatives of (B1) can be computed by substituting (28) into it and then noting that it can be written as

$$\text{Prob } (D_i|T, r_{c_i}) = K_1 e^{-(\alpha+\beta) r_{c_i} v} + K_2 e^{-\beta r_{c_i} v} \left(1 - e^{-\alpha r_{c_i} v}\right) + K_3 \left(1 - e^{-\beta r_{c_i} v}\right), \tag{B2}$$

which is easily rearranged into

$$\text{Prob } (D_i|T, r_{c_i}) = (K_1 - K_2) e^{-(\alpha+\beta) r_{c_i} v} + (K_2 - K_3) e^{-\beta r_{c_i} v} + K_3, \tag{B3}$$

where

$$K_1 = \sum_x \pi_x \ell^{(i)}_{jc_i}(x) \ell^{(i)}_{kc_i}(x), \tag{B4}$$

$$K_2 = \sum_x \pi_x \ell^{(i)}_{jc_i}(x) \sum_y \left(\frac{\pi_x}{\sum_k \pi_y \epsilon_{xy}} \right) \epsilon_{xy} \ell^{(i)}_{kc_i}(y), \tag{B5}$$

and

$$K_3 = \left(\sum_x \pi_x \ell^{(i)}_{jc_i}(x) \right) \left(\sum_y \pi_y \ell^{(i)}_{kc_i}(y) \right), \tag{B6}$$

so that the derivatives are simply

$$\frac{d \text{ Prob } (D_i|T, r_{c_i})}{dv} = -r_{c_i}(\alpha+\beta)(K_1 - K_2) e^{-(\alpha+\beta) r_{c_i} v} - r_{c_i} \beta (K_2 - K_3) e^{-\beta r_{c_i} v} \tag{B7}$$

and

$$\frac{d^2 \text{ Prob } (D_i|T, r_{c_i})}{dv^2} = r^2_{c_i}(\alpha+\beta)^2 (K_1 - K_2) e^{-(\alpha+\beta) r_{c_i} v} + r^2_{c_i} \beta^2 (K_2 - K_3) e^{-\beta r_{c_i} v}. \tag{B8}$$

Literature Cited

Baldi, P., Y. Chauvin, T. Hunkapiller, and M. A. McClure. 1994. Hidden Markov models of biological primary sequence information. Proc. Nat. Acad. Sci. USA **91**:1059-1063.

Baum, L. E. and T. Petrie. 1966. Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Stat. **37**:1554-1563.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss. 1970. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Stat. **41**:164-171.

Churchill, G. A. 1989. Stochastic models for heterogeneous DNA sequences. Bull. Math. Biol. **51**:79-94.

Felsenstein, J. 1973. Maximum-likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. **22**:240-249.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368-376.

Forney, G. D, Jr. 1973. The Viterbi algorithm. Proc. IEEE **61**:268-278.

Hasegawa, M., H. Kishino, and T. Yano. 1988. Phylogenetic inference from DNA sequence data. Pp. 1-13 *in* K. Matusita, ed. Statistical Theory and Data Analysis II: Proceedings of the Second Pacific Area Statistical Conference, North-Holland, Amsterdam.

Haussler, D., A. Krogh, I. S. Mian and K. Sjolander. 1993. Protein modeling using hidden Markov models: analysis of globins. Pp. 792-802 *in* T. N. Mudge, V. Milutinovic and L. Hunter, eds. Proceeding of the Twenty-Sixth Hawaii International Conference on System Sciences. IEEE, Los Alamitos, California.

Hein, J. 1993. A heuristic method to reconstruct the history of sequences subject to

recombination. J. Mol. Evol. **36**:369-405.

Higgins, D. G. and P. M. Sharp. 1989. Fast and sensitive multiple sequence alignments on a microcomputer. Comput. Appl. Biosci. **5**:151-153.

Jin, L. and M. Nei. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. **7**:82-102.

Jukes, T. H. and C. Cantor. 1969. Evolution of protein molecules. Pp. 21-132 *in* M. N. Munro, ed. Mammalian Protein Metabolism. Academic Press, New York.

Kelly, C. and J. Rice. 1995. Modeling nucleotide evolution: a heterogeneous rate analysis. Math. Biosci., in press.

Kimura, M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111-120.

Kishino, H. and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. **29**:170-179.

Kitagawa, G. 1987. Non-Gaussian state-space modelling of nonstationary time series. J. Am. Stat. Assoc. **82**:1032-1041.

Krogh, A., M. Brown, I. S. Mian, K. Sjölander, and D. Haussler. 1994. Hidden Markov models in computational biology. Applications to protein modeling. J. Mol. Biol. **235**:1501-1531.

Neyman, J. 1971. Molecular studies of evolution: a source of novel statistical problems. Pp. 1-27 *in* S. S. Gupta and J. Yackel, eds. Statistical Decision Theory and Related Topics. New York: Academic Press.

Olsen, G. J. 1987. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harbor Symp. Quant. Biol. **52**:825-837.

Rempe, U. 1988. Characterizing DNA variability by stochastic matrices. Pp. 375-384 *in* H. H. Bock, ed. Classification and Related Methods of Data Analysis. Proceedings of the First Conference of the International Federation of Classication Societies (IFCS), Technical University of Aachen, F.R.G, 29 June-1 July, 1987. North-Holland, Amsterdam.

Swofford, D. L. and G. J. Olsen. 1990. Phylogeny reconstruction. Chapter 11, Pp. 411-501 *in* D. M. Hillis and C. Moritz, eds. Molecular Systematics. Sinauer Associates, Sunderland, Massachusetts.

Tillier, E. R. M. 1994. Maximum likelihood with multiparameter models of substitution. J. Mol. Evol. **39**:409-417.

Tillier, E. R. M. and R. A. Collins. 1995. Neighbor-joining and maximum likelihood with RNA sequences: addressing the inter-dependence of sites. Mol. Biol. Evol. **12**:7-15.

Viterbi, A. J. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans. Inf. Theory **IT-13**:260-269.

von Haeseler, A. and G. A. Churchill. 1993. Network models for sequence evolution. J. Mol. Evol. **37**:77-85.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10**:1396-1401.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306-314.

Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. Genetics, in press.

Figure Captions

Fig. 1. A representation of the model used in this paper. The phylogeny for the species is shown to the left of the sequences, the potential and actual hidden states for each site are shown below them.

Fig. 2. The phylogeny estimated for the eight hemoglobin β DNA sequences. The shorter branches are not statistically significant.

Fig. 3. The β-hemoglobin coding sequences used in the data example. The dots are sites at which the sequence is the same as in *Tachyglossus*. The two rows of digits below each section of sequences are the regional rate categories inferred for each site. The first shows the single combination of regional rate assignments that contributes most the the likelihood. The second shows an assignment for each site provided that 95% or more the likelihood is contributed by that rate being assigned to that site (otherwise no assignment is shown). Category 1 has the lower rate.

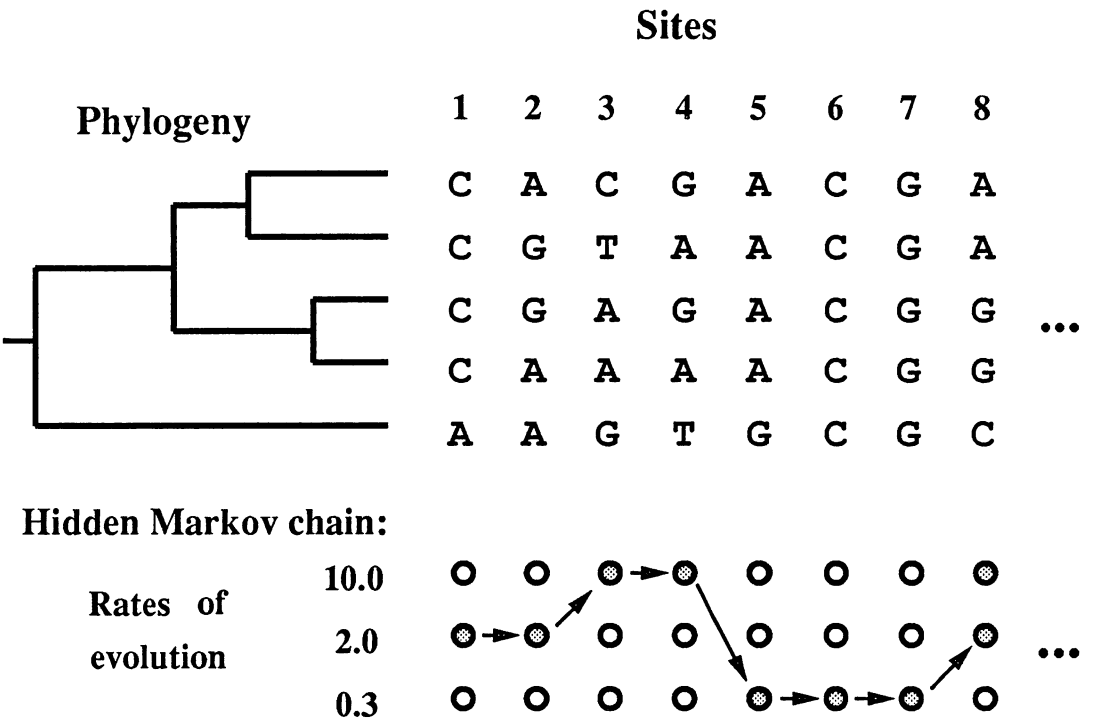
Sites
Phylogeny
1 2 3 4 5 6 7 8
C A C G A C G A
C G T A A C G A
C G A G A C G G ...
C A A A A C G G
A A G T G C G C
Hidden Markov chain:
Rates of evolution
10.0
2.0
0.3
...

Figure 1

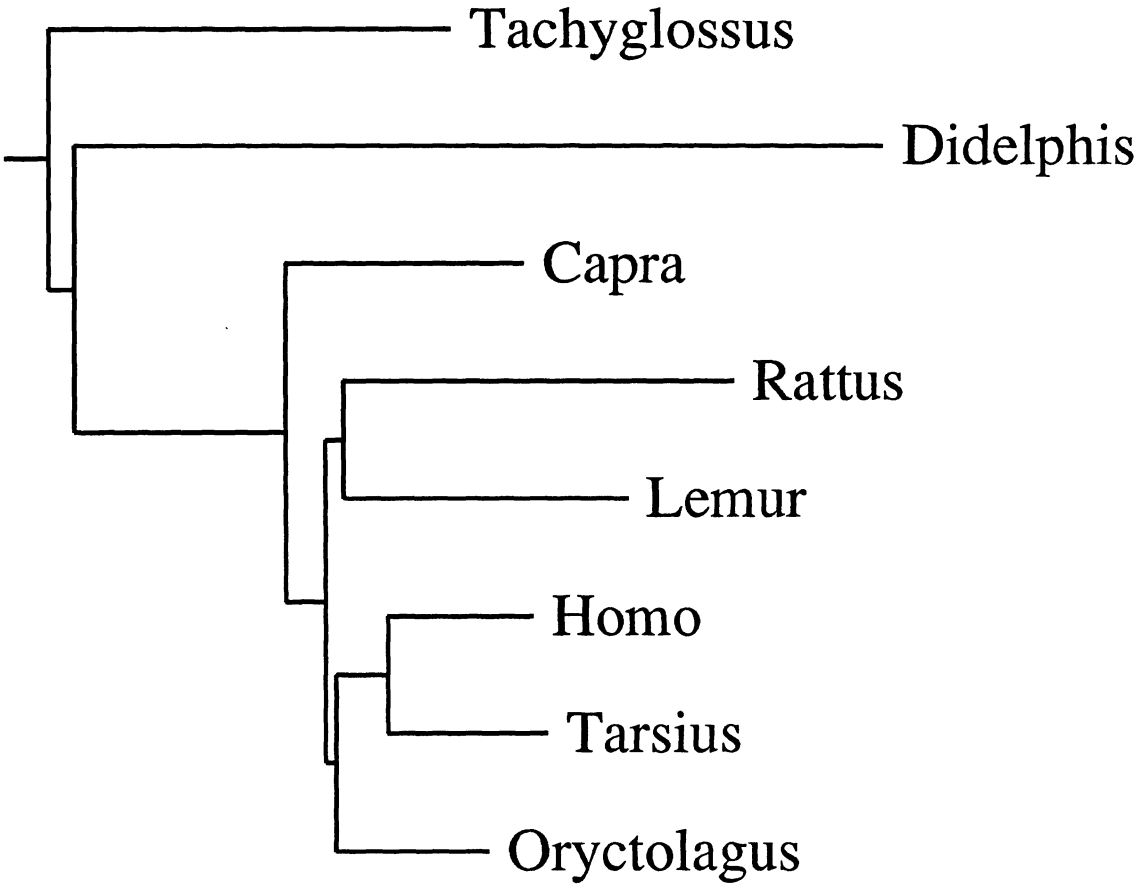
Tachyglossus
Didelphis
Capra
Rattus
Lemur
Homo
Tarsius
Oryctolagus

Figure 2