Combining Phylogenetic and Hidden Markov Models in Biosequence Analysis

Adam Siepel
Center for Biomolecular Science and Engr.
University of California
Santa Cruz, CA 95064, USA
acs@soe.ucsc.edu

David Haussler
Howard Hughes Medical Institute and
Center for Biomolecular Science and Engr.
University of California
Santa Cruz, CA 95064, USA
haussler@soe.ucsc.edu

ABSTRACT

A few models have appeared in recent years that consider not only the way substitutions occur through evolutionary history at each site of a genome, but also the way the process changes from one site to the next. These models combine phylogenetic models of molecular evolution, which apply to individual sites, and hidden Markov models, which allow for changes from site to site. Besides improving the realism of ordinary phylogenetic models, they are potentially very powerful tools for inference and prediction—for gene finding, for example, or prediction of secondary structure. In this paper, we review progress on combined phylogenetic and hidden Markov models and present some extensions to previous work. Our main result is a simple and efficient method for accommodating higher-order states in the HMM, which allows for context-sensitive models of substitution—that is, models that consider the effects of neighboring bases on the pattern of substitution. We present experimental results indicating that higher-order states, autocorrelated rates, and multiple functional categories all lead to significant improvements in the fit of a combined phylogenetic and hidden Markov model, with the effect of higher-order states being particularly pronounced.

General Terms

Algorithms, Experimentation, Theory

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and genetics; I.6.8 [**Simulation and Modeling**]: Types of Simulation—*Combined*; G.3 [**Probability and Statistics**]: Markov processes

Keywords

Maximum likelihood, context-sensitive substitution

RECOMB'03, April 10–13, 2003, Berlin, Germany.

1. INTRODUCTION

Since their introduction to bioinformatics about a decade ago [5, 22, 23], hidden Markov models (HMMs) have become one of the dominant tools in biological sequence analysis. They are especially important in the areas of gene prediction [24, 4, 21] and homology searching [22, 8, 20], but have been applied to a wide variety of problems [5, 7, 37]. While fundamentally more powerful models, such as stochastic context free grammars, are preferable for certain applications [6], HMMs appear in many cases to strike the right balance between simplicity and expressiveness.

An important limitation of most HMMs in use today, however, is that they fail to take advantage of the best available models of sequence evolution. More than three decades of research have produced powerful, probabilistic models of evolution that consider not only the topology of the phylogenetic tree by which present-day sequences are related, but also the lengths of its branches, and the pattern of substitution by which the sequences have evolved [9, 32, 10, 11, 41, 39]. As phylogenetic models of molecular evolution and HMMs are both explicitly probabilistic (and indeed, are applied using very similar algorithms), it would seem that the best aspects of both might be incorporated into a single framework.

HMMs generally work *along the length* of a sequence, mostly ignoring the evolutionary process at each site; phylogenetic models work *across sequences*, more or less ignoring changes along their length. To borrow a notion from Yang [43], HMMs (as applied to biological sequences) operate in the dimension of *space* and phylogenetic models in the dimension of *time*. Because they are orthogonal in this way, the two types of models turn out to be fairly easy to combine. Combined phylogenetic and hidden Markov models were derived independently by Felsenstein and Churchill [13] and Yang [43] to allow for autocorrelation of evolutionary rates at different sites, with the goal of improving the realism of models of evolution and the accuracy of phylogenetic inferences. Subsequently, Thorne, Goldman, Jones, and Lío [37, 14, 26] applied a very similar type of combined model to the problem of secondary structure prediction, recognizing that the basic paradigm could be useful for various types of comparative sequence analysis. (Somewhat different combinations of HMMs and phylogeny have also been applied to the problems of multiple alignment [17] and combined tree-building/multiple alignment [28]). Despite these efforts, however, combined models remain little used, and have yet to be applied to many suitable problems.

We believe that the combined phylogenetic and hidden Markov model could become an important general-purpose tool in the bioinformatician's arsenal. These combined models, in a sense, are the natural extension of HMMs for the comparative genomics era: they allow the information contained in multiple, aligned sequences to be brought to bear on the problems of spatial discrimination for which HMMs are so effective, and in doing so they remain purely probabilistic, interpretable with efficient algorithms, and reasonably faithful to the underlying biology. We think it especially important that they explicitly use the phylogeny—which not only represents the evolutionary relationships of the sequences in question, but, as Goldman *et al.* have pointed out [14], also defines their correlation structure.

In this paper, we review and extend previous work on combined phylogenetic and hidden Markov models. The paper begins with an overview of phylogenetic models, including extensions that allow rate variation among sites. Next, we review HMMs designed to allow autocorrelation of evolutionary rate, and then present a simple extension that accommodates *functional categories* as well as *rate categories.* Following this, we introduce a simple and efficient method for computing the emission probabilities of *higher-order states.* This method allows for *neighbor-* or *context-sensitive* models of base substitution, which consider the N bases preceding each base, and are capable of capturing the dependency of substitution patterns on neighboring bases [3, 29]. Finally, we present the results of a small experimental study indicating that higher-order states, autocorrelated rates, and multiple functional categories all dramatically improve the fit of the model, and the improvements are roughly additive. The effect of higher order states (context-sensitivity) is particularly pronounced. We have focused here on the question of how best to improve the fit of a combined model; applying such a model to problems of inference and prediction remains a subject for future work.

2. METHODS

We assume a correct multiple alignment of n sequences of length L, with one sequence for each of n taxa. We further assume that the taxa are related by a phylogenetic tree of known topology, and we begin with standard simplifying assumptions about the substitution process: it is homogeneous throughout the tree, and it acts independently and identically at different columns of the alignment (we will partly relax the latter assumption later in the paper). Let us denote the alignment as $\mathbf{X} = \{x_{i,j}\}$, with $x_{i,j}$ being the jth character in the ith sequence ($1 \leq i \leq n, 1 \leq j \leq L$). For now, we assume that every $x_{i,j}$ is drawn from an alphabet $\mathbf{\Sigma}$ (this assumption will also be relaxed). In this paper, we will use $\mathbf{\Sigma} = $ (A,C,G,T) (for computational purposes, it is convenient to impose an ordering on the alphabet), but the methods apply to any alphabet, provided an appropriate substitution model is available (see below). The jth column of the alignment will be denoted $\mathbf{X}_j$. We will use the terms "column" and "site" interchangeably.

Let a *tree model* $\boldsymbol{\psi}$ be defined as a tuple of four parameters, $\boldsymbol{\psi} = (\mathbf{Q}, \boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\pi})$, with $\mathbf{Q}$ a substitution rate matrix, $\boldsymbol{\tau}$ a tree topology, $\boldsymbol{\beta}$ a vector of branch lengths, and $\boldsymbol{\pi}$ a vector of equilibrium base frequencies. The matrix $\mathbf{Q}$ is of dimension $|\mathbf{\Sigma}| \times |\mathbf{\Sigma}|$, and defines a continuous-time Markov process for base substitution (to be detailed below). The topology $\boldsymbol{\tau}$, in general, is a binary tree with n leaves, and thus has $2n-1$ nodes and $2n-2$ edges (usually called "branches"). For some substitution models, however (ones known as "reversible"), the tree will be unrooted, and thus will have one fewer node and edge. The vector $\boldsymbol{\beta}$ assigns a non-negative real value to each branch of the tree, representing its evolutionary length, usually as an expected number of substitutions per site (see below). The parameters $\boldsymbol{\tau}$ and $\boldsymbol{\beta}$ can be described in such a way that there is an implicit mapping between leaves of the tree and rows of the alignment, and between edges of the tree and elements of $\boldsymbol{\beta}$ (details are not important here). The vector $\boldsymbol{\pi}$ is of dimension $|\mathbf{\Sigma}|$ and describes the background frequency at which each base appears. We will adopt the commonly used practice of estimating $\boldsymbol{\pi}$ directly from $\mathbf{X}$, by measuring the observed frequency of each base, and subsequently considering it a fixed parameter (but see [38]).

2.1 Computing the Likelihood of a Tree Model

The critical step of computing the likelihood of a given tree model, $P(\mathbf{X}|\boldsymbol{\psi})$, was worked out more than twenty years ago [32, 10, 11]. The "pruning algorithm" of Felsenstein uses dynamic programming and the assumption of site-independence to render feasible the imposing task of summing the likelihoods of all possible labelings of ancestral nodes of a tree. The assumption of independence allows $P(\mathbf{X}|\boldsymbol{\psi}) = \prod_{i=1}^{L} P(\mathbf{X}_i|\boldsymbol{\psi})$ and reduces the problem to that of computing the likelihood of each column $\mathbf{X}_i$. However, $P(\mathbf{X}_i|\boldsymbol{\psi}) = \sum_{\mathcal{L}} P(\mathcal{L}, \mathbf{X}_i|\boldsymbol{\psi})$, where $\mathcal{L}$ is a labeling of the $n-1$ ancestral nodes of the tree with elements from $\mathbf{\Sigma}$ (the labels at the leaves are fixed by $\mathbf{X}_i$); this sum has $O(|\mathbf{\Sigma}|^n)$ terms. Felsenstein's algorithm works as follows. Let u be any node in $\boldsymbol{\tau}$ and let v and w be its children (if u is a leaf, then v and w are null). In addition, let t_v and t_w be the lengths of the branches connecting u to v and u to w, respectively (if u is a leaf, then t_v and t_w are also null). Suppose for the moment that we can compute the probability of base b replacing base a over a branch of length t, which we will denote $P(b|a,t)$. Now, following Durbin *et al.* [6], we denote by $P(L_u|a)$ the probability of all of the leaves below node u given that the base assigned node u is an a (implicitly conditioned on $\boldsymbol{\psi}$. The algorithm says that

$$P(L_u|a) = \begin{cases} I(a = x_u) & \text{if } u \text{ is a leaf} \\ \displaystyle\sum_b P(b|a,t_v)P(L_v|b) \times \\ \qquad \displaystyle\sum_c P(c|a,t_w)P(L_w|c) & \text{otherwise} \end{cases} \tag{1}$$

where I is the indicator function and x_u is the element of $\mathbf{X}_i$ corresponding to leaf u. Thus, for each base a and node u, the likelihood of the leaves beneath u can be computed directly from the likelihoods of the leaves beneath its children, considering all possible labels at each child. The likelihoods at the leaves are trivially 1 or 0, depending on the observed data. To find the total likelihood of the column, we consider each possible label at the root r of the tree, weighted by its equilibrium frequency; that is, $P(\mathbf{X}_i|\boldsymbol{\psi}) = \sum_a \pi_a P(L_r|a)$. A post-order traversal of the tree takes $O(n)$ time (where n is the number of species), the computation at each node is essentially constant (equation 1), and the entire procedure must be repeated for each of the L columns in the alignment. Thus, the complete computation of $P(\mathbf{X}|\boldsymbol{\psi})$ is linear in the size of the alignment.

The maximum-likelihood tree model is defined as

$$\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} P(\mathbf{X}|\boldsymbol{\psi}). \tag{2}$$

Estimation of $\hat{\boldsymbol{\psi}}$ is usually accomplished by partitioning the free parameters of $\boldsymbol{\psi}$ into $\boldsymbol{\tau}$ and $(\mathbf{Q}, \boldsymbol{\beta})$; for any given tree topology, $(\mathbf{Q}, \boldsymbol{\beta})$ are optimized using a quasi-Newton or EM algorithm, and this step is repeated for all possible values of $\boldsymbol{\tau}$ (*a priori* knowledge or heuristic methods may be used to reduce the set of topologies to consider). If the topology is assumed to be known, as in our case, the problem is greatly simplified.

2.2 Models of DNA Substitution

Felsenstein's algorithm depends on efficient computation of $P(b|a, t)$, the probability that a base b is substituted for a base a over a branch of length t, for any bases $a, b \in \mathbf{\Sigma}$ and any non-negative real value t. Such probabilities are generally based on a continuous-time Markov model of base substitution, with the instantaneous rate of replacement of each base for each other defined by the rate matrix $\mathbf{Q}$ [46, 39]. The various available substitution models can be seen as alternative ways of parameterizing $\mathbf{Q}$, usually with the goal of reducing as much as possible the number of free parameters to optimize while still providing a sufficiently rich model. As a continuous-time Markov matrix, $\mathbf{Q} = \{q_{i,j}\}$ $(1 \le i, j \le |\mathbf{\Sigma}|)$ is constrained to have each of its rows sum to zero; furthermore, the matrix is by convention scaled so that the expected rate of substitution at equilibrium is one, which has the effect of establishing the unit of branch lengths to be *expected substitutions per site*. Thus, for $|\mathbf{\Sigma}| = 4$ (we will limit our discussion here to DNA sequences), $\mathbf{Q}$ has at most $4^2 - 4 - 1 = 11$ free parameters. Allowing all 11 parameters to be free results in the "unrestricted" (UNREST or UNR) substitution model. Imposing the constraint of "reversibility," which says that $\pi_i q_{i,j} = \pi_j q_{j,i}$ for all i and j, reduces the number of parameters to 5 (the REV model). One of the simplest models still regarded as reasonably realistic is that of Hasegawa, Kishino, and Yano [16] (the HKY model), which has a single parameter, κ, representing the ratio of the rates of transitions to transversions. The UNR, REV, and HKY substitution models, all of which will be used in this study, correspond to parameterizations of $\mathbf{Q}$ as follows[1]:

$$\mathbf{Q}_{\text{UNR}} = \begin{pmatrix} - & a & b & c \\ d & - & e & f \\ g & h & - & i \\ j & k & l & - \end{pmatrix} \quad \mathbf{Q}_{\text{REV}} = \begin{pmatrix} - & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & - & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & - & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & - \end{pmatrix}$$

$$\mathbf{Q}_{\text{HKY}} = \begin{pmatrix} - & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & - & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & - \end{pmatrix}$$

The $-$ symbols along the main diagonals indicate elements to be defined as $q_{i,i} = -\sum_j q_{i,j} I(i \ne j)$.

The probability of any substitution along a branch of length t is obtainable as a function of $\mathbf{Q}$ and t. Let $\mathbf{P}(t)$ be the matrix of substitution probabilities for length t (note that $\mathbf{P}(t)$ is a *discrete* Markov matrix, with rows summing to 1, while $\mathbf{Q}$ is a *continuous* Markov matrix, with rows summing to 0). $\mathbf{P}(t)$ is given by the solution to the differential equation $\frac{d}{dt}\mathbf{P}(t) = \mathbf{P}(t)\mathbf{Q}$ with initial conditions $\mathbf{P}(0) = \mathbf{I}$, which is $\mathbf{P}(t) = e^{\mathbf{Q}t}$. $\mathbf{Q}$ is generally diagonalizable as $\mathbf{Q} = \mathbf{S\Lambda S}^{-1}$, allowing $\mathbf{P}(t)$ to be computed as $\mathbf{P}(t) = \mathbf{S}e^{\mathbf{\Lambda}t}\mathbf{S}^{-1}$, where $e^{\mathbf{\Lambda}t}$ is the diagonal matrix obtained by exponentiating each element on the main diagonal of $\mathbf{\Lambda}t$.

2.3 Allowing for Different Rates at Different Sites

An obvious deficiency of the original method of Felsenstein is its assumption that evolution occurs at each site by an identical process. Yang has partly addressed this problem with an elegant extension of the method that allows variation in the *rate* of evolution at different sites. His method effectively factors out of the branch-length vector $\boldsymbol{\beta}$ a parameter r, which is assumed to be a random variable having a gamma distribution, $r \sim \text{Gamma}(\alpha, \beta)$ (p.d.f. $g(r; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r},\ r > 0$). Likelihoods are computed by integrating over all possible values of r, either accounting for a fully continuous distribution [41], or using a discrete approximation [42]. The shape parameter α is estimated from the data; the scale parameter β is simply set equal to α, ensuring a mean of 1. The discrete approximation partitions the distribution into k rate categories of equal probability (the distribution is "cut" at quantiles $\frac{1}{k}, \frac{2}{k}, \dots, \frac{k-1}{k}$), and defines a rate constant r_j $(1 \le j \le k)$ for each category, such that r_j is equal to the mean (or alternatively, the median) of the jth partition. The probability $P(\mathbf{X}_i|\boldsymbol{\psi})$ can be approximated as

$$P(\mathbf{X}_i|\boldsymbol{\psi}) = \sum_{j=1}^{k} \frac{1}{k} \cdot P(\mathbf{X}_i|(\mathbf{Q}, \boldsymbol{\tau}, r_j\boldsymbol{\beta}, \boldsymbol{\pi})). \tag{3}$$

This value can be computed with k invocations of Felsenstein's algorithm, and thus only increases the cost of the likelihood computation by a factor of k. Yang showed that the improvement in likelihoods drops off quickly after about $k = 3$ and recommended $k = 4$ as an appropriate choice. Notice that $k = 1$ corresponds to Felsenstein's original single-rate model, and $k = \infty$ to the continuous gamma model.

Felsenstein and Churchill [13], observing that any "halfway realistic" model of rate variation should also reflect the tendency of evolutionary pressures to act in similar ways at spatially proximate positions, introduced a model that also uses a discrete set of rate categories, but additionally assumes sites are assigned categories by a Markov process, which is defined by an autocorrelation parameter λ. In their model, if column $\mathbf{X}_{i-1}$ is assigned category j, then with probability λ, column $\mathbf{X}_i$ will be assigned category j, and with probability $1 - \lambda$, $\mathbf{X}_i$ will be assigned a category drawn at random from the equilibrium distribution for all categories, denoted $\mathbf{f}$. Thus, the transition probabilities between the k modes are given by a $k \times k$ Markov matrix $\mathbf{C} = \{c_{j,l}\}$, with

$$c_{j,l} = \lambda I(j = l) + (1 - \lambda) f_l. \tag{4}$$

This Markov process will achieve the distribution $\mathbf{f}$ at stationarity. Felsenstein and Churchill showed that the total probability of an alignment (which must consider all possible assignments of categories) can be computed efficiently using a recursive dynamic-programming algorithm, which is equivalent to the forward algorithm [6]. Similarly, the maximum-likelihood assignment of categories can be obtained by the Viterbi algorithm, and the posterior probability that each site is assigned each category can be obtained by posterior decoding (see Section 2.4). Yang independently developed a very similar method, which uses the rate categories defined by the discrete gamma method, and derives the transition

[1] Note that one degree of freedom is lost in each case due to the constraint on the scaling of the matrix—that is, that $\sum_{i,j} \pi_i q_{i,j} I(i \ne j) = 1$.

probabilities between them according to a bivariate gamma distribution. Thus, the parameter α, which is a free parameter in the fitting procedure, defines the rate categories (Felsenstein and Churchill defined them manually). Another free parameter, ρ, fills the role of λ, but influences the transition probabilities in a more complex way.

In this paper, we use a hybrid strategy: we define rate categories according to the discrete gamma method, but we define autocorrelation in terms of the parameter λ, using equation 4. In this way, the rate categories are chosen to fit the data, but we avoid some complexity in coding and simplify the extension to multiple functional categories. This method also allows us to use a uniform distribution for $\mathbf{f}$ (see equation 4), because of the way the rate categories are chosen; thus, $c_{j,l} = \frac{1-\lambda}{k}$ if $j \neq l$ and $c_{j,j} = \lambda + \frac{1-\lambda}{k}$. We consider λ as a free parameter, but simplify the fitting process by proceeding in two steps: first, we estimate α, along with $\mathbf{Q}$ and $\boldsymbol{\beta}$ (using to the standard discrete gamma method), then we estimate λ with all other parameters fixed (a one-dimensional line search is adequate here; we use Brent's method [34]). This strategy is not guaranteed to find the true maximum-likelihood estimate of $(\boldsymbol{\psi}, \alpha, \lambda)$, but it seems to provide a close approximation, because the other parameters are insensitive to λ (as noted by Yang [43]). An approximate method is important in our case, because we must accommodate multiple independent tree models and HMMs with potentially many states (see below). Our model reduces to the discrete gamma model when $\lambda = 0$.

Finally, note that our definition of a tree model can be extended to allow for rate variation by including the parameter α; that is, $\boldsymbol{\psi} = (\mathbf{Q}, \boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\pi}, \alpha)$ (we will consider λ separately). For the remainder of the paper, assume this extension.

2.4 Allowing for Different Categories of Sites

The idea of modeling changes in rate as a Markov process can be generalized to allow for k arbitrary tree models, $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k$, not necessarily related in any particular way. The only requirement is that we have estimates of the probability that any site $\mathbf{X}_i$ obeys model $\boldsymbol{\psi}_j$, given that the previous site, $\mathbf{X}_{i-1}$ obeys model $\boldsymbol{\psi}_{j'}$, which together constitute a $k \times k$ Markov matrix. The different tree models might describe different "functional categories" of sites, rather than different rate categories. For example, one category might consist of sites in 1st codon positions, others of sites in 2nd and 3rd codon positions, and another of non-coding sites; the Markov transition probabilities might reflect the high probability that a 2nd codon position follows a 1st codon position, the low but non-zero probability that a non-coding site follows a 2nd codon position, and the zero probability that a 3rd codon position follows a 1st codon position. Alternatively, the functional categories might correspond to secondary structural characteristics, as in the models of Thorne, Goldman, *et al.* [37, 14, 26], or to any biological property that changes in some non-random way along the length of biological sequences (e.g., GC content, CpG incidence, tertiary structure). Usually, the topologies of the tree models will be the same, but no such constraint is required; indeed, allowing different topologies could be useful in detecting hybridization, horizontal transfer, or viral recombination.

Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_i, \ldots, \phi_L)$, $1 \leq \phi_i \leq k$, denote an assignment of tree models to sites, representing the hypothesis that each site obeys the assigned tree model (ϕ_i is the index of the tree model assigned to column $\mathbf{X}_i$). In addition, let $\mathbf{A} = \{a_{j,l}\}$ $(1 \leq j, l \leq k)$ be the $k \times k$ matrix of transition probabilities between tree models. The joint probability of the entire alignment $\mathbf{X}$ and the assignment $\boldsymbol{\phi}$ is given by

$$P(\mathbf{X}, \boldsymbol{\phi}) = \prod_{i=1}^{L} a_{\phi_{i-1},\phi_i} P(\mathbf{X}_i | \boldsymbol{\psi}_{\phi_i}) \qquad (5)$$

where $\phi_0 = 0$ and a_{0,ϕ_1} is the probability of beginning with tree model ϕ_1. Described in this way, the model can be seen to be a hidden Markov model whose emission probabilities are determined according to a phylogenetic model of molecular evolution (the Markov chain for state transitions is homogeneous and first-order). In practice, the dimensions of "space" (the transition probabilities of the HMM) and "time" (the phylogenetic model) are quite separable: one needs only to compute $P(\mathbf{X}_i|\boldsymbol{\psi}_j)$ for all $1 \leq i \leq L$ and all $1 \leq j \leq k$, then to pass these values to generic HMM routines as an $L \times k$ matrix of emission probabilities. The standard algorithms are available to compute the total probability of $\mathbf{X}$ ($\sum_{\boldsymbol{\phi}} P(\mathbf{X}, \boldsymbol{\phi})$; the forward algorithm), a maximum-likelihood assignment ($\arg\max_{\boldsymbol{\phi}} P(\mathbf{X}, \boldsymbol{\phi})$; the Viterbi algorithm), and the posterior probability that any site $\mathbf{X}_i$ is assigned tree model j ($P(\phi_i = j|\mathbf{X})$; posterior decoding). See Durbin *et al.* [6] for an accessible but thorough discussion of these algorithms.

It is common in phylogenetic analysis to label the columns of an alignment with functional categories, and to compute the probability of each column conditional on its label. This technique can result in dramatically improved likelihoods, due to quite different substitution properties at sites of different types [44]. By modeling functional categories with an HMM, however, we can benefit in the same way even when sequences are unannotated or annotation is unreliable. Furthermore, this framework provides a means to *infer* or *predict* the correct label at each site—which may be taken, for example, to be the one corresponding to a maximum-likelihood assignment, or the one of highest posterior probability. Note that we have allowed differences between functional categories in the *proportions* of branch lengths, or in the *pattern* of substitution (as described by the rate matrix), as well as in the rate of evolution. This capability will be especially important when inferring functional labels, because it provides an additional means for discriminating between sites of different categories.

In this paper, we assume a "supervised learning" strategy for the inference or prediction of labels, with a training step based on labeled data. A tree model can be fit separately to each functional category of the labeled training data, and the HMM transition probabilities can be estimated by a simple counting method (using pseudocounts as necessary to avoid overfitting). It would be possible, however, to learn HMM transitions directly from the unlabeled data, using the Baum-Welch algorithm [6]. As long as the algorithm is initialized with approximately the right parameter settings, it should converge on a reasonable HMM. Such a strategy would be expensive, however, as all tree models would need to be re-estimated on each iteration of the algorithm.

Having argued that rate categories are subsumed by something more general, we now distinguish them again as a special case. Even in sites of the same functional category, evolutionary rate appears to vary at the regional level, as well as from site to site [27, 40, 30]. Rate categories, therefore, are in a sense orthogonal to functional categories. It may be especially useful when inferring functional categories to allow changes in rate *within* each functional class. For example, we need to be able to distinguish a slow-evolving non-coding region from a medium- or fast-evolving coding region. Both functional categories and rate categories can be accommodated if we create several "scaled" versions of each functionally determined tree model, and define the Markov transition matrix as a cross product of two HMMs: one with a state for each rate category, and one with a state for each functional category (the implicit assumption is that the two Markov processes are independent).

In particular, suppose we have k rate categories and q functional categories. Suppose further that the q functional categories are described by distinct tree models $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_q$, that the transition probabilities among the functional categories are given by a $q \times q$ matrix $\mathbf{F} = \{f_{i,j}\}$ $(1 \leq i, j \leq q)$, and that the transition probabilities among the rate categories are given by a $k \times k$ matrix $\mathbf{C} = \{c_{i',j'}\}$ $(1 \leq i', j' \leq k$; $\mathbf{C}$ may be defined by equation 4). Let $r_{i,i'}$ $(1 \leq i \leq q,\ 1 \leq i' \leq k)$ be the rate constant for the i'th category of $\boldsymbol{\psi}_i$. We define a new sequence of kq tree models, $(\boldsymbol{\psi}'_{1,1}, \ldots, \boldsymbol{\psi}'_{1,k}, \ldots, \boldsymbol{\psi}'_{q,1}, \ldots, \boldsymbol{\psi}'_{q,k})$, such that, for all $1 \leq i \leq q$ and $1 \leq i' \leq k$,

$$\boldsymbol{\psi}'_{i,i'} = r_{i,i'}\boldsymbol{\psi}_i = (\mathbf{Q}_i, \boldsymbol{\tau}_i, r_{i,i'}\boldsymbol{\beta}_i, \boldsymbol{\pi}_i, \alpha_i) \qquad (6)$$

In addition, we define a new $kq \times kq$ transition matrix $\mathbf{A} = \{a_{l,m}\}$ $(1 \leq l, m \leq kq)$ such that, for all $1 \leq i, j \leq q$, and $1 \leq i', j' \leq k$,

$$a_{(i-1)k+i',(j-1)k+j'} = f_{i,j}c_{i',j'}. \qquad (7)$$

The methods described above can now be applied without change to $(\boldsymbol{\psi}'_{1,1}, \ldots, \boldsymbol{\psi}'_{q,k})$ and $\mathbf{A}$ with the desired effect. It may be necessary when interpreting results, however, to "project" states onto the dimension of interest; for example, the "raw" Viterbi path might be expressed in terms of functional categories only.

2.5 Allowing for Missing Data

With real data, it is usually the case that certain characters in the alignment $\mathbf{X}$ are not consistent with the assumed evolutionary process, in that they do not belong to the alphabet $\boldsymbol{\Sigma}$. Alignment gaps are the most common source of such characters, but they may also result from polymorphism within species, or from failure of the sequencing process to resolve bases unambiguously. It is common in phylogenetic analysis simply to discard any column containing a character not in $\boldsymbol{\Sigma}$; this practice, however, is undesirable for alignments of divergent sequences, in which only a small minority of columns may be completely without gaps. (Another strategy sometimes used for gaps, which has obvious deficiencies, is to treat the gap character as an additional element in $\boldsymbol{\Sigma}$). Various alternatives have been proposed for handling gaps, including ones that actually derive phylogenetic information from them [28]. In this paper, we employ a simple approach in which gaps and all other characters not in $\boldsymbol{\Sigma}$ are regarded uniformly as missing data. This method is essentially neutral with respect to such characters; they neither contribute phylogenetic information nor take away from it by "contaminating" a portion of the alignment. The method is not novel—indeed, it was briefly mentioned in Felsenstein's 1973 paper [10] and has been implemented in PHYLIP [12] and PAML [45], among other packages—but we will describe it in some detail because it turns out to be useful in the extension of Section 2.6.

Consider a single column of an alignment, $\mathbf{X}_i$, some elements of which are missing. Let M be the set of all columns that result from assigning characters from the alphabet $\boldsymbol{\Sigma}$ in place of missing elements in $\mathbf{X}_i$. The total probability of $\mathbf{X}_i$ is the sum of the probabilities of all elements of M, $P(\mathbf{X}_i|\boldsymbol{\psi}) = \sum_{\mathbf{Y}\in M} P(\mathbf{Y}|\boldsymbol{\psi})$. (If all elements of $\mathbf{X}_i$ are missing, then $P(\mathbf{X}_i|\boldsymbol{\psi}) = 1$). It may be helpful to regard the missing elements of $\mathbf{X}_i$ as "wildcards" and to denote them "*". M can be thought of as the set of columns that "match" $\mathbf{X}_i$, allowing for wildcards. For example, a column $\mathbf{X}_i$ =(A,C,C,*,A)$^{\mathrm{T}}$ has $M = \{$(A,C,C,A,A)$^{\mathrm{T}}$, (A,C,C,C,A)$^{\mathrm{T}}$, (A,C,C,G,A)$^{\mathrm{T}}$, (A,C,C,T,A)$^{\mathrm{T}}\}$. Notice that incomplete wildcards are also possible. For example, the ambiguity character R (purine) might be allowed to match only A or G.

Felsenstein's algorithm, because it is *already* summing over possible assignments of characters to nodes in the tree, requires only a very minor change to accommodate missing data of this kind. Recall that the base case of the recursion, which is applied when a node u is a leaf, is $P(L_u|a) = I(a = x_u)$, for any base a (see equation 1). To allow for missing data, we need only replace "$a = x_u$" with "a matches x_u." Thus, at a leaf u corresponding to a "*", $P(L_u|a) = 1$ for all a, and as the algorithm works its way from the leaves to the root of the tree, all possible assignments of bases to u will be considered. With missing data allowed, equation 1 generalizes to

$$P(L_u|a) = \begin{cases} I(a \text{ matches } x_u) & \text{if } u \text{ is a leaf} \\ \displaystyle\sum_b P(b|a,t_v)P(L_v|b)\ \times \\ \quad \displaystyle\sum_c P(c|a,t_w)P(L_w|c) & \text{otherwise} \end{cases} \qquad (8)$$

Thus, despite that the set M may be exponentially large, missing data can be accommodated with no additional cost in computation.

2.6 An Extension to Higher-Order States

Our description so far has assumed so-called "0th order" states, in which the emission probability for column $\mathbf{X}_i$ at state (tree model) j, $P(\mathbf{X}_i|\boldsymbol{\psi}_j)$, depends only on column $\mathbf{X}_i$. Much additional discriminatory power can be gained in many biological applications through the use of higher-order states. In an Nth order state, the emission probability of $\mathbf{X}_i$ at state j is conditioned on the previous N columns, $\mathbf{X}_{i-N}, \ldots, \mathbf{X}_{i-1}$. (Gene finders may have states with N as large as 4 or 5 [21, 4]). In this section, we show how to extend the methods discussed so far to the case of $N > 0$. The essential problem is to compute $P(\mathbf{X}_i|\mathbf{X}_{i-1}, \ldots, \mathbf{X}_{i-N})$ (here and in the discussion below, conditioning on $\boldsymbol{\psi}_j$ is implicit). By considering conditional emission probabilities of this type, we will effectively model substitution as a neighbor- or context-sensitive process.

Felsenstein's algorithm can readily be adapted to compute the *joint* probability, $P(\mathbf{X}_{i-N}, \ldots, \mathbf{X}_i)$: simply assume an alphabet of size $|\boldsymbol{\Sigma}|^{N+1}$, consisting of all $(N+1)$-tuples of characters in $\boldsymbol{\Sigma}$, and adjust the dimensions of $\mathbf{Q}$ (the rate matrix) and $\boldsymbol{\pi}$ (the vector of equilibrium frequencies) accordingly. (This is essentially what Goldman and Yang have done for their codon-based model [15]). In the case of $N = 1$, corresponding to dinucleotides, Felsenstein's algorithm can be written

$$P(L_u|a_1a_2) = \begin{cases} I(a_1a_2 \text{ matches } x_{u,1}x_{u,2} & \text{if } u \text{ is a leaf} \\ \displaystyle\sum_{b_1b_2} P(b_1b_2|a_1a_2,t_v)P(L_v|b_1b_2) \times & \\ \quad \displaystyle\sum_{c_1c_2} P(c_1c_2|a_1a_2,t_w)P(L_w|c_1c_2) & \text{otherwise} \end{cases} \tag{9}$$

where a_1a_2, b_1b_2, and c_1c_2 are variables representing dinucleotides, $x_{u,1}x_{u,2}$ is the dinucleotide corresponding to leaf u, and the definition of "matching" is extended appropriately for dinucleotides. The computation of the terms of the form $P(b_1b_2|a_1a_2,t)$ works exactly as before, except that $\mathbf{Q}$ and $\mathbf{P}(t)$ are now of dimension $|\mathbf{\Sigma}|^2$. Similarly, $P(\mathbf{X}_{i-1},\mathbf{X}_i)$ can be obtained as before from $\boldsymbol{\pi}$ and the values $P(L_r|a_1a_2)$ at the root r, but now $\boldsymbol{\pi}$ is of dimension $|\mathbf{\Sigma}|^2$.

The problem is that we need *conditional* rather than joint probabilities. We can use the property that

$$P(\mathbf{X}_i|\mathbf{X}_{i-N},\ldots,\mathbf{X}_{i-1}) = \frac{P(\mathbf{X}_{i-N},\ldots,\mathbf{X}_i)}{\sum_{\mathbf{Y}} P(\mathbf{X}_{i-N},\ldots,\mathbf{X}_{i-1},\mathbf{Y})}$$

where $\sum_{\mathbf{Y}}$ is the sum over all 4^n possible assignments of bases in $\mathbf{\Sigma}$ to the ith column in the alignment. Despite its exponential size, this sum can be computed efficiently using dynamic programming (as might be expected). It turns out we have already solved the problem: it can be regarded as an instance of the missing data problem. The sum $\sum_{\mathbf{Y}} P(\mathbf{X}_{i-N},\ldots,\mathbf{X}_{i-1},\mathbf{Y})$ is the same as $P(\mathbf{X}_{i-N},\ldots,\mathbf{X}_{i-1},\mathbf{Z})$, with $\mathbf{Z} = (*,*,\ldots,*)^{\mathrm{T}}$, by our definition of missing data. Thus, $P(\mathbf{X}_i|\mathbf{X}_{i-N},\ldots,\mathbf{X}_{i-1})$ can be computed by two passes through Felsenstein's algorithm, with a different initialization for each pass[2]. The time to compute the likelihood of an entire alignment is in general $O(nL|\mathbf{\Sigma}|^{N+1})$, for n sequences and an alignment of length L. In practice, N must remain small (probably at most 2 or 3), for there to be sufficient data to estimate $\mathbf{Q}$, and for the number of free parameters to be kept manageable. For the remainder of this paper, we will assume $N = 0$ (single nucleotides) or $N = 1$ (dinucleotides).

Our method as a whole, then, can be summarized as follows. Assume k rate categories, q functional categories, and an HMM with states of order N. Also assume that the matrix $\mathbf{F}$ of transition probabilities between functional categories and tree models $\boldsymbol{\psi}_1,\ldots,\boldsymbol{\psi}_q$ have been estimated from labeled training data. To apply the model to an unlabeled data set, create a new set of kq tree models, according to equation 6, and compute the $kq \times L$ matrix of emission probabilities, consisting of $P(\mathbf{X}_i|\boldsymbol{\psi}_j)$ for all $1 \leq i \leq L$ and $1 \leq j \leq kq$. Use the missing-data version of Felsenstein's algorithm (equation 8), generalized if necessary for $N > 0$, as described above. Now, starting, with an arbitrary value for the parameter λ (e.g., $\lambda = 0.9$), construct a matrix $\mathbf{C}$ according to equation 4, and define $\mathbf{A}$ as the cross product of $\mathbf{F}$ and $\mathbf{C}$, according to equation 7. Compute the total likelihood, using the forward algorithm. Repeat the final steps until a value of λ is found that maximizes the likelihood (in the case of $k = 1$, no iteration is necessary). If the Viterbi path or posterior probabilities are desired, obtain them in a final pass, with λ fixed at its MLE. Notice that, with our approximate method for estimating λ, the emission probabilities need only be computed once.

For the case of dinucleotides, it remains to find a suitable way to parameterize the 16×16 rate matrix $\mathbf{Q}$. We consider three alternatives: a fully reversible dinucleotide matrix (R2), a strand-symmetric reversible matrix (R2S), and a strand-symmetric unrestricted matrix (U2S). In all cases, to reduce the number of free parameters, we prohibit instantaneous changes involving more than one base (despite biological evidence for such changes [2]; as future work, we intend to relax this restriction). Thus, R2 is defined with $\mathbf{Q} = \{q_{i_1i_2,j_1j_2}\}$ $(1 \leq i_1,i_2,j_1,j_2 \leq |\mathbf{\Sigma}|)$ such that:

$$q_{i_1i_2,j_1j_2} = \begin{cases} 0 & \text{if } i_1 \neq j_1 \text{ and } i_2 \neq j_2 \\ a_{i_1i_2,j_1j_2}\pi_{j_1j_2} & \text{else if } j_1j_2 > i_1i_2 \\ a_{j_1j_2,i_1i_2}\pi_{j_1j_2} & \text{else if } j_1j_2 < i_1i_2 \\ -\sum_{k_1k_2 \neq j_1j_2} q_{i_1i_2,k_1k_2} & \text{else } (i_1i_2 = j_1j_2) \end{cases}$$

where $j_1j_2 < i_1i_2$ means $(j_1-1)m+j_2 < (i_1-1)m+i_2$, and the parameters of the form $a_{i_1i_2,j_1j_2}$ are all free (there are 48). R2S is identical to R2, except for the constraint that $a_{i_1i_2,j_1j_2} = a_{i'_1i'_2,j'_1j'_2}$ if $i'_1i'_2$ is the reverse complement of i_1i_2 and $j'_1j'_2$ is the reverse complement of j_1j_2, which cuts the number of parameters in half to 24. U2S is identical to R2S except reversibility is not required; it has 48 free parameters.

2.7 Implementation and Data

Code was written in C to support both training and testing. Training is accomplished directly from sequence annotations, with transition probabilities between functional categories estimated by counting (pseudocounts optional), and tree models estimated separately for subsets of alignment sites corresponding to each of the specified functional categories. Optimization of parameters is accomplished using a quasi-Newton algorithm (BFGS [34]), with gradients computed using the difference method. The code was shown to give equivalent results to those of the PAML package [45] for shared models. The core routine to compute the likelihood of a tree model supports missing data and higher-order states, as well as the discrete gamma method (code to compute gamma quantiles and mean rates was borrowed from PAML). The HKY, REV, UNR, R2, R2S, and U2S substitution models were all implemented, as were the forward, backward, and Viterbi algorithms. The software is available upon request[3].

For both training and testing, we used portions of a multiple alignment consisting of nearly two megabases of human sequence, in the region of the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene (chromosome 7), and homologous sequence from 8 other eutherian mammals [36] (see species in Figure 1). The sequences are products of the NISC Comparative Sequencing Program[4]. The multiple alignment, which is derived from pairwise local alignments, was created by Webb Miller using the MultiPipMaker program[5]. It contains 2.3 million columns, 1.9 million of which correspond to human bases. Each column was labeled with one of seven functional categories (codon positions 1, 2, and 3, intron, $5'$ UTR, $3'$ UTR, and intergenic),

[2] Some care is required when $i \leq N$. One can simply regard the missing columns as missing data, and apply the same principle once more. For example, with $N = 1$, $P(\mathbf{X}_1|\mathbf{X}_0) = \frac{P(\mathbf{Z},\mathbf{X}_1)}{P(\mathbf{Z},\mathbf{Z})} = P(\mathbf{Z},\mathbf{X}_1)$, where $\mathbf{Z} = (*,\ldots,*)^{\mathrm{T}}$.

[3] An updated version is under development and intended to be made available at `http://www.cse.ucsc.edu/∼acs`.

[4] `http://www.nisc.nih.gov`

[5] MultiPipMaker is a recent extension to PipMaker [35]; see `http://bio.cse.psu.edu/pipmaker/`

using annotations prepared at the NISC via a combination of computational and manual techniques (Pamela Jacques Thomas, personal correspondence). Gene structure in this region is almost perfectly conserved across species [36]. The annotations include 10 genes, which together cover 21,084 of the human sequence's 1.9 million bases, or about 1% (excluding UTRs). The sequenced region is generally AT rich (GC content 38.5% in human and 38.6% over all species).

We have focused our analysis on two subsets of sites in the alignment: one corresponding to "ancestral repeats" (ARs), believed to reflect neutral evolution, and another corresponding to one of the genes, WNT2, selected because it is relatively short in total length, yet contains significant representation from all functional classes. The AR subset consisted of sites corresponding to segments of the human sequence that had been identified by the RepeatMasker program[6] as belonging to repeat families that are believed to be ancestral to eutherian mammals—that is, dispersed and rendered quiescent prior to the eutherian radiation. The same set of families was used as described recently by the Mouse Genome Sequencing Consortium [30]. More such sites were available than we could efficiently analyze with our current software, so we selected 20,000 consecutive sites belonging to L1 transposons, with good representation across species. The WNT2 alignment was constructed by simply extracting the segment of the whole alignment corresponding to the WNT2 gene, along with 2,000 bases of intergenic DNA on either side. About 60,000 columns resulted of which 50,062 contained human sequence. 1,083 (2.2%) of these corresponded to coding regions (361 in each codon position), distributed fairly evenly among five exons. No sequence was available in this region for two of the nine species, pig and dog.

The phylogenetic tree relating the nine species was assumed to have the (unrooted) topology shown in Figure 1. This topology is consistent with the accepted taxonomy of the species, and does appear to have the highest likelihood under reversible models (by a wide margin). For the nonreversible UNR model, we rooted the tree on the branch separating the rodents from the other species, as resulted in the highest likelihood. Note, however, that the true root appears to group the rodents and the primates to the exclusion of the artiodactyls and carnivores [31, 36].

3. RESULTS

We first discuss the effect of using models of increasing richness for base substitution and for rate variation, within sites of a single functional category. Where appropriate, we compare models using the standard likelihood ratio test (LRT), which is based on the assumption that twice the difference in their log likelihoods obeys a χ^2 distribution with d degrees of freedom, where d is equal to the difference in the number of free parameters of the models [18]. Figure 2 shows log likelihoods for the AR alignment under the five different substitution models and under three models for rate variation (constant rates, the discrete gamma model with $k = 4$, and the autocorrelation model, also with $k = 4$). The likelihoods are seen steadily to improve as the models become richer in terms of both substitution and rate variation; however, the improvement obtained from replacing a single-nucleotide substitution model with a dinucleotide model (e.g., REV with R2) overshadows all others by an order of magnitude. The LRT is only applicable among REV, HKY, and UNR, which are "nested" (UNR subsumes REV which subsumes HKY), and among R2S, U2S, and R2 (R2S is subsumed by both U2S and R2). It is also applicable across rate models for nested substitution models. Among nested models, the improvement of each model over the next simpler alternative is easily statistically significant, except in the case of R2 over R2S[7]. It appears that the simplifying assumption of strand symmetry is reasonable in neutral DNA; the assumption of reversibility is less well-supported. The LRT is not applicable between the single-nucleotide and dinucleotide models, but the improvement of the dinucleotide models appears to be overwhelmingly significant

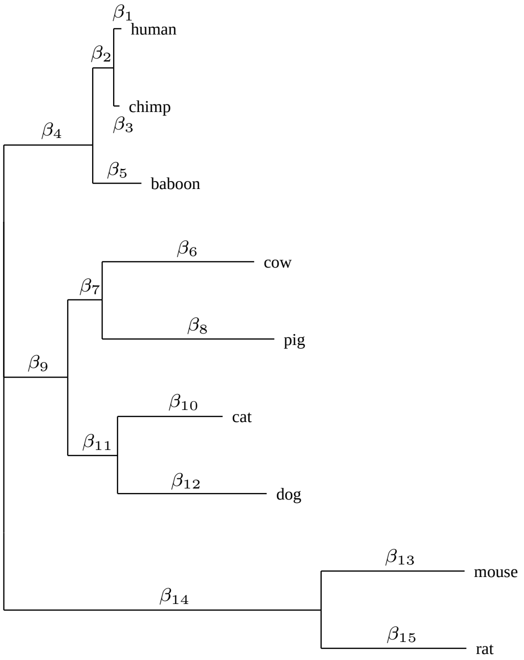

Figure 1: Phylogenetic tree assumed for the nine species (unrooted). Branch lengths are drawn in the proportions estimated for the AR alignment using the REV model (with discrete gamma). Values for the branch lengths as estimated under various models are presented in a supplementary appendix.

Estimates of key parameters were generally similar under all models. The branch lengths were highly consistent, although they did exhibit a tendency (as has been noted previously [46]) to increase slightly with models of increasing richness (see supplementary appendix[8]). Estimates of the parameter α, which determines the shape of the gamma distribution in models that allow rate variation, were similar under single-nucleotide models (5.73 [HKY], 5.96 [REV], and 6.06 [UNR]), but interestingly, increased significantly when switching to dinucleotide substitution models (8.60

[6] `http://ftp.genome.washington.edu/cgi-bin/RepeatMasker/`

[7] The LRT shows R2 to offer a significant improvement over R2S with constant rates, but in the discrete gamma and autocorrelated cases, the improvement is marginal.

[8] Available at `http://www.cse.ucsc.edu/~acs/pubs.html`.

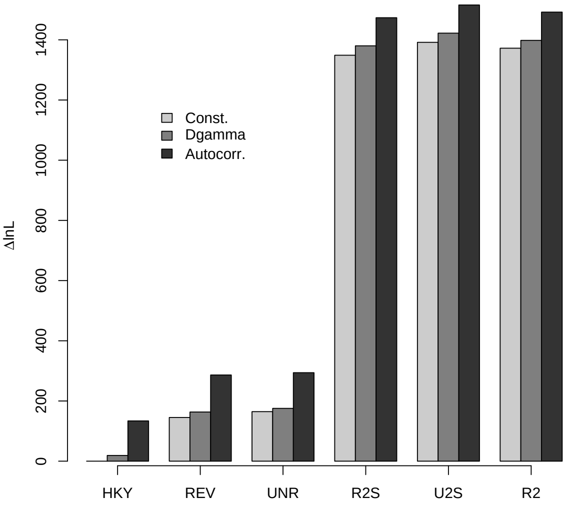

Figure 2: Log likelihoods for the AR alignment under various substitution and rate models. Values are shown relative to HKY with constant rates.

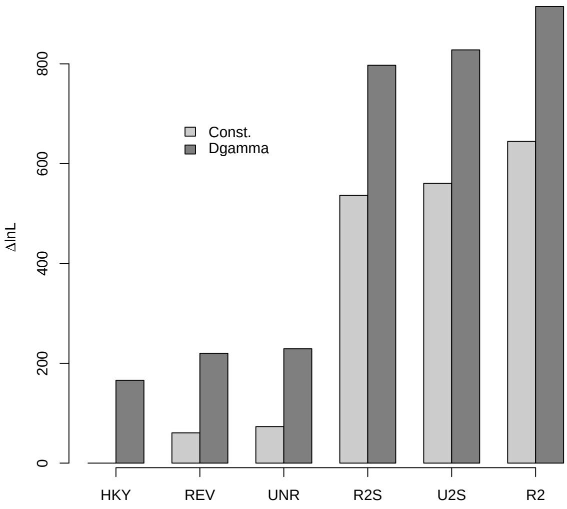

Figure 3: Log likelihoods for sites in the second codon position, under various substitution and rate models. Values are shown relative to HKY with constant rates.

[R2S], 8.73 [U2S], and 9.27 [R2])[9]. Evidently, some apparent rate variation can be explained by considering the context of each substitution. The autocorrelation parameter λ was fairly insensitive to the substitution model (all estimates were between 0.95 and 0.97). For the dinucleotide models, the 16×16 rate matrix $\mathbf{Q}$ reflected a very strong "CpG effect" (high mutation rate of CG to TG, due to methylation and spontaneous deamination [25]); indeed, the estimates of the rate of change from CG to TG (and its reverse complement, CA) exceeded all others by nearly an order of magnitude (see supplementary appendix). The other differences in substitution rates appear to be mostly explainable in terms of the transition/transversion bias.

To contrast coding and neutral DNA, we performed a similar experiment with all bases in the 2nd codon position of the entire 2-megabase alignment (Figure 3) (we did not apply the autocorrelation model to this data set, because it consists of sites that are not adjacent). The results were similar, except that the advantage of the dinucleotide models (which here describe bases in the 1st and 2nd codon positions) is much less pronounced, and the improvement due to the discrete gamma model is somewhat more pronounced (as expected). Not surprisingly, strand symmetry is seen to be a far poorer assumption in coding DNA (compare the differences between R2S and R2 in Figures 2 and 3). The estimated rate matrix (not shown) appears to capture something about the pattern of amino acid substitutions, but the effect is imperfect, because many dinucleotide substitutions correspond to mixtures of quite different amino acid substitutions. This is apparently what causes the dinucleotide models to be of less benefit in coding than in noncoding regions. Nevertheless, many rate-matrix parameters appeared to be informative. For example under R2 with constant rates, after adjusting for equilibrium frequencies, one of the largest rates was for GT↔AT (Val↔Ile/Met, BLOSUM62 scores 3 and 1), and one of the smallest for AT↔AA (Ile/Met↔Asn/Lys, BLOSUM62 scores −3, −3, −2, and −1).

Next, we examine the effect of considering functional categories. Figure 4 shows log likelihoods for the WNT2 alignment, with 1, 4, and 6 functional categories, and various models for substitution and rate variation. For the 4-category model, we classified all sites as 1st, 2nd, or 3rd codon positions, or "other", and for the the 6-category model, we partitioned the "other" category into introns, $5'$ UTR sites, and intergenic sites (separating out $3'$ UTR sites seems to be of little benefit). The transition probabilities of each HMM were estimated from the WNT2 alignment itself, by counting observed transitions; no pseudocounts were used (we chose here more or less to avoid issues of training, so that we might focus on other effects). In the case of multiple functional categories and autocorrelated rates, a cross-product HMM was used, as discussed in Section 2.4. A very large improvement is seen in moving both from 1-category to 4-category models and from 4-category to 6-category models under both REV and R2 (note the scale of the graph), with the improvements under R2 somewhat more pronounced. For each substitution model and category combination, improvements achieved by moving to the discrete gamma model are significant but modest; however, enormous improvements are realized by introducing autocorrelation. This effect appears to result from highly autocorrelated rates (estimates of λ

[9] Estimates of α were quite high in all cases, consistent with the assumption of unrestricted, neutral evolution. The low rate variation in this data set explains the relatively minor improvement seen with the discrete gamma model.

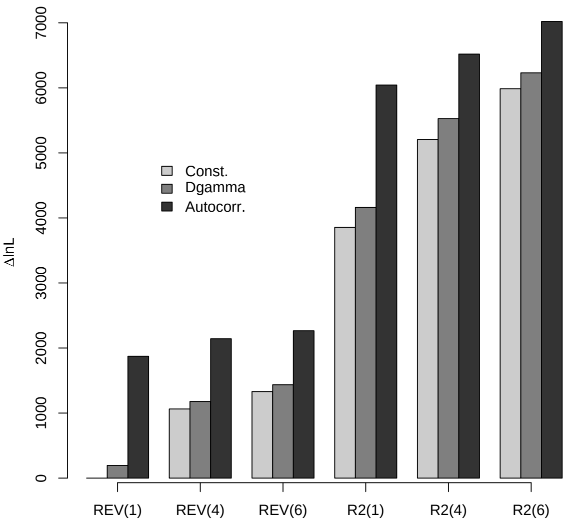

Figure 4: Log likelihoods for the WNT2 alignment, for the REV and R2 substitution models, three models of rate variation, and sets of 1, 4, and 6 functional categories. Values are shown relative to REV with constant rates and one functional category.

were all between 0.991 and 0.994)[10]. Comparison with likelihoods conditional on *a priori* labels of functional categories (which in this case are the same ones used to train the HMM and tree models), indicate that the HMMs result in an increase of between 220 and 490 units of log likelihood (results not shown).

4. DISCUSSION

A hidden Markov model and a phylogenetic model can be combined to create a new model of molecular evolution that captures both spatial and temporal aspects of the process. Our results suggest that the goodness of fit of such a model is improved significantly by allowing for context-sensitive substitution rates (corresponding to higher-order states), autocorrelated rate variation, and multiple functional categories. Furthermore, each type of improvement occurs more or less independently of the others, and their combination is especially powerful. Judging by the case of dinucleotides, context-sensitive substitution models are particularly beneficial, especially in neutral DNA, where the CpG effect is strong. The improvement in coding regions is somewhat muted by a mismatch with the naturally occurring "word size" of three. Because of the CpG effect, a word size of two may turn out to be optimal for noncoding DNA; for coding DNA, however, three is almost certain to be the magic number. Our software is not yet up to the task of nucleotide triples (which require many more free parameters and larger data sets, as well as a small increase in the cost of each likelihood evaluation), but we expect they can be accommodated with some tuning of the code.

It should be noted that our solution to the problem of context-sensitivity is only approximate, in that it does not allow for overlapping substitutions of $(N + 1)$-tuples along any single branch in the tree. A fully general model would allow context-sensitive effects to cascade to both sides of a given $(N + 1)$-tuple. Such a model requires that the Markov chain underlying the continuous-time Markov process for base substitution be replaced with a Markov random field. This problem has recently been addressed for the two-sequence case [19, 33, 1], but extensions to complete phylogenies appear computationally intensive (without appropriate approximations), and remain as future work.

We have only begun to apply combined hidden Markov and phylogenetic models to problems of inference and prediction, but we are optimistic about their potential. One very straightforward application of the model is the identification of conserved regions in an alignment. Some preliminary experiments on this problem, which used the posterior probabilities of columns being assigned to rate categories, have been encouraging.

5. ACKNOWLEDGEMENTS

We thank Eric Green for permission to use the sequence data prior to its publication, Webb Miller for his invaluable work on multiple alignment, Arian Smit for identifying the ancestral repeat families, and Ryan Weber for determining their locations in the data set. Nick Goldman and Simon Whelan provided helpful comments on the dinucleotide model, and Ziheng Yang's PAML package has been useful in many ways throughout this project. Ziheng Yang and an anonymous reviewer are also acknowledged for alerting us to previous work on the missing data problem. This work was supported by the Howard Hughes Medical Institute (D.H.) and NHGRI grant IP41HG02371 (A.S.).

6. REFERENCES

[1] P. Arndt, C. Burge, and T. Hwa. DNA sequence evolution with neighbor-dependent mutation. In G. Myers, S. Hannenhalli, S. Istrail, P. Pevzner, and M. Waterman, editors, *Proceedings of the Sixth Annual International Conference on Computational Biology*, pages 32–38, New York, 2002. ACM.

[2] M. Averof, A. Rokas, K. H. Wolfe, and P. Sharp. Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science*, 287:1283–1286, 2000.

[3] R. Blake, S. Hess, and J. Nicholson-Tuell. The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *J. Mol. Evol.*, 34:189–200, 1992.

[4] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268:78–94, 1997.

[5] G. Churchill. Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.*, 51:79–94, 1989.

[6] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[7] S. Eddy. Multiple alignment using hidden Markov models. In C. Rawlings, D. Clark, R. Altman, L. Hunter, T. Lengauer, and S. Wodak, editors, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 114–120. AAAI Press, 1995.

[10] Models with multiple functional categories have not resulted in higher estimates of λ, as might have been expected. The reason may be that, for the 1-state models, strong autocorrelation in non-coding regions overwhelms weak autocorrelation in coding regions (recall that the latter make up only about 2% of all sites).

[8] S. Eddy. Profile hidden Markov models. *Bioinformatics*, 14:755–763, 1998.

[9] A. Edwards and L. Cavalli-Sforza. Reconstruction of evolutionary trees. In V. Heywood and J. McNeill, editors, *Phenetic and phylogenetic classification*, pages 67–76. Systematics Association, London, 1964.

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

[11] J. Felsenstein. Evolutionary trees from DNA sequences. *J. Mol. Evol.*, 17:368–376, 1981.

[12] J. Felsenstein. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle, 1993. Available from `http://evolution.genetics.washington.edu/phylip.html`.

[13] J. Felsenstein and G. Churchill. A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, 13:93–104, 1996.

[14] N. Goldman, J. Thorne, and D. Jones. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol.*, 263:196–208, 1996.

[15] N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, 11:725–735, 1994.

[16] M. Hasegawa, H. Kishino, and T. Yano. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 22:160–174, 1985.

[17] I. Holmes and W. Bruno. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, 17(9):803–820, 2001.

[18] J. Huelsenbeck and B. Rannala. Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. *Science*, 276:227–232, 1997.

[19] J. Jensen and A.-M. Pedersen. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv. Appl. Prob*, 32:499–517, 2000.

[20] K. Karplus, C. Barrett, and R. Hughey. Hidden Markov models for detecting remote protein homologs. *Bioinformatics*, 14:846–856, 1998.

[21] A. Krogh. Two methods for improving performance of an HMM and their application for gene finding. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, pages 179–186. AAAI Press, 1997.

[22] A. Krogh, M. Brown, I. Mian, K. Sjölander, and D. Haussler. Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.*, 235:1501–1531, 1994.

[23] A. Krogh, I. Mian, and D. Haussler. A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic Acids Res.*, 22:4768–4778, 1994.

[24] D. Kulp, D. Haussler, M. Reese, and F. Eeckman. A generalized hidden Markov model for the recognition of human genes in DNA. In D. States, P. Agarwal, T. Gaasterland, L. Hunter, and R. Smith, editors, *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, pages 134–142, Menlo Park, CA, 1996. AAAI Press.

[25] B. Lewin. *Genes VII*. Oxford University Press, 2000.

[26] P. Lió, N. Goldman, J. Thorne, and D. Jones. PASSML: combining evolutionary inference and protein secondary structure prediction. *Bioinformatics*, 14:726–733, 1998.

[27] G. Matassi, P. Sharp, and C. Gautier. Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.*, 9:786–791, 1999.

[28] G. Mitchison. A probabilistic treatment of phylogeny and sequence alignment. *J. Mol. Evol.*, 49(1):11–22, 1999.

[29] B. Morton, V. Oberholzer, and M. Clegg. The influence of specific neighboring bases on substitution bias in noncoding regions of the plant chloroplast genome. *J. Mol. Evol.*, 45:227–231, 1997.

[30] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.

[31] W. Murphy et al. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science*, 294:2348–2351, 2001.

[32] J. Neyman. Molecular studies of evolution: A source of novel statistical problems. In S. Gupta and J. Yackel, editors, *Statistical Decision Theory and Related Topics*, pages 1–27. Academic Press, New York, 1971.

[33] A.-M. Pedersen and J. Jensen. A dependent rates model and MCMC based methodology for the maximum likelihood analysis of sequences with overlapping reading frames. *Mol. Biol. Evol.*, 18:763–776, 2001.

[34] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, second edition, 1992.

[35] S. Schwartz, Z. Zhang, K. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. PipMaker: A web server for aligning two genomic DNA sequences. *Genome Res.*, 10(4):577–586, 2000.

[36] J. Thomas, J. Touchman, R. Blakesley, G. Bouffard, S. Beckstrom-Sternberg, E. Margulies, M. Blanchette, A. Siepel, P. Thomas, J. McDowell, B. Maskeri, N. Hansen, M. Schwartz, R. Weber, W. Kent, D. Karolchik, T. Bruen, R. Bevan, D. Cutler, S. Schwartz, L. Elnitski, J. Idol, A. Prasad, S.-Q. Lee-Lin, V. Maduro, M. Portnoy, N. Dietrich, N. Akhter, K. Ayele, B. Benjamin, K. Cariaga, C. Brinkley, S. Brooks, S. Granite, X. Guan, J. Gupta, P. Haghighi, S.-L. Ho, M. Huang, E. Karlins, P. Laric, R. Legaspi, M. Lim, Q. Maduro, C. Masiello, S. Mastrian, J. McCloskey, R. Pearson, S. Stantripop, E. Tiongson, J. Tran, C. Tsurgeon, J. Vogt, M. Walker, K. Wetherby, L. Wiggins, A. Young, L.-H. Zhang, K. Osoegawa, B. Zhu, B. Zhao, C. Shu, P. D. Jong, C. Lawrence, A. Smit, A. Chakravarti, D. Haussler, P. Green, W. Miller, and E. Green. Multi-species sequencing of targeted genomic regions: Comparative studies of genome structure, function, and evolution. In preparation, 2003.

[37] J. Thorne, N. Goldman, and D. Jones. Combining protein evolution and secondary structure. *Mol. Biol. Evol.*, 13:666–673, 1996.

[38] S. Whelan and N. Goldman. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol. Biol. Evol.*, 16(9):1292–1299, 1999.

[39] S. Whelan, P. Liò, and N. Goldman. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet.*, 17(5):262–272, 2001.

[40] E. Williams and L. Hurst. The proteins of linked genes evolve at similar rates. *Nature*, 407:900–903, 2000.

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

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

[43] Z. Yang. A space-time process model for the evolution of DNA sequences. *Genetics*, 139:993–1005, 1995.

[44] Z. Yang. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.*, 42:587–596, 1996.

[45] Z. Yang. PAML: A program package for phylogenetic analysis by maximum likelihood. *CABIOS*, 13:555–556, 1997.

[46] Z. Yang, N. Goldman, and A. Friday. Comparison of models for nucleotide substution used in maximum likelihood phylogenetic estimation. *Mol. Biol. Evol.*, 11(2):316–224, 1994.

APPENDIX

	HKY	**REV**	**UNR**	**R2S**	**U2S**	**R2**
β_1	0.0073	0.0073	0.0073	0.0075	0.0075	0.0073
β_2	0.0195	0.0193	0.0192	0.0202	0.0203	0.0196
β_3	0.0058	0.0058	0.0058	0.0060	0.0060	0.0057
β_4	0.0778	0.0788	0.0788	0.0841	0.0870	0.0817
β_5	0.0433	0.0435	0.0435	0.0451	0.0451	0.0435
β_6	0.1327	0.1336	0.1331	0.1423	0.1415	0.1380
β_7	0.0307	0.0310	0.0309	0.0328	0.0322	0.0319
β_8	0.1495	0.1505	0.1500	0.1593	0.1584	0.1545
β_9	0.0558	0.0554	0.0549	0.0602	0.0568	0.0583
β_{10}	0.0926	0.0927	0.0927	0.0994	0.0994	0.0964
β_{11}	0.0439	0.0439	0.0437	0.0477	0.0471	0.0462
β_{12}	0.1306	0.1309	0.1305	0.1366	0.1357	0.1320
β_{13}	0.1245	0.1249	0.1248	0.1327	0.1324	0.1282
β_{14}	0.2672	0.2672	0.2680	0.2871	0.2866	0.2777
β_{15}	0.1280	0.1271	0.1280	0.1336	0.1345	0.1292

Table 1: Branch lengths for the tree shown in Figure 1, as estimated for the AR alignment under various substitution models. The values shown assume constant rates; values allowing for rate variation are similar but tend to be slightly larger.

	Equilibrium Frequencies															
	0.09	0.05	0.06	0.10	0.07	0.04	0.01	0.06	0.05	0.03	0.03	0.06	0.09	0.06	0.07	0.14
	Q															
AA	−1.31	0.13	0.46	0.12	0.16	0.00	0.00	0.00	0.28	0.00	0.00	0.00	0.13	0.00	0.00	0.00
AC	0.22	−2.04	0.18	0.86	0.00	0.16	0.00	0.00	0.00	0.42	0.00	0.00	0.00	0.16	0.00	0.00
AG	0.72	0.14	−1.69	0.18	0.00	0.00	0.17	0.00	0.00	0.00	0.35	0.00	0.00	0.00	0.11	0.00
AT	0.13	0.50	0.14	−1.69	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.55	0.00	0.00	0.00	0.19
CA	0.16	0.00	0.00	0.00	−2.08	0.21	0.66	0.14	0.11	0.00	0.00	0.00	0.79	0.00	0.00	0.00
CC	0.00	0.20	0.00	0.00	0.18	−2.66	0.20	0.94	0.00	0.13	0.00	0.00	0.00	0.99	0.00	0.00
CG	0.00	0.00	1.10	0.00	**9.02**	0.88	−22.25	1.24	0.00	0.00	0.71	0.00	0.00	0.00	**9.27**	0.00
CT	0.00	0.00	0.00	0.18	0.10	0.43	0.17	−2.11	0.00	0.00	0.00	0.16	0.00	0.00	0.00	1.04
GA	0.58	0.00	0.00	0.00	0.16	0.00	0.00	0.00	−1.74	0.09	0.57	0.11	0.20	0.00	0.00	0.00
GC	0.00	0.66	0.00	0.00	0.00	0.22	0.00	0.00	0.27	−2.40	0.18	0.73	0.00	0.32	0.00	0.00
GG	0.00	0.00	0.83	0.00	0.00	0.00	0.20	0.00	0.83	0.13	−2.41	0.22	0.00	0.00	0.18	0.00
GT	0.00	0.00	0.00	0.86	0.00	0.00	0.00	0.21	0.14	0.42	0.13	−2.10	0.00	0.00	0.00	0.32
TA	0.15	0.00	0.00	0.00	0.68	0.00	0.00	0.00	0.13	0.00	0.00	0.00	−2.07	0.16	0.70	0.22
TC	0.00	0.10	0.00	0.00	0.00	0.70	0.00	0.00	0.00	0.09	0.00	0.00	0.20	−2.13	0.16	0.85
TG	0.00	0.00	0.13	0.00	0.00	0.00	0.66	0.00	0.00	0.00	0.17	0.00	0.79	0.13	−2.12	0.23
TT	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.52	0.00	0.00	0.00	0.15	0.13	0.33	0.17	−1.44
	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT

Table 2: Rate matrix Q estimated for the AR alignment, under the U2S substitution model, along with estimated dinucleotide equilibrium frequencies. The value at row i and column j, indicates the instantaneous rate at which dinucleotide j substitutes for dinucleotide i. The values corresponding to the the CpG effect are highlighted.