JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 11, Numbers 2–3, 2004

Pp. 413–428

Combining Phylogenetic and Hidden Markov Models in Biosequence Analysis

ADAM SIEPEL[1] and DAVID HAUSSLER[1,2]

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 example, for gene finding 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-dependent 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.

Key words: phylogenetic models, hidden Markov models, gene prediction, maximum likelihood, context-dependent substitution.

1. INTRODUCTION

SINCE THEIR INTRODUCTION TO BIOINFORMATICS about a decade ago (Churchill, 1989; Krogh *et al.*, 1994b, 1994a), 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 (Kulp *et al.*, 1996; Burge and Karlin, 1997; Krogh, 1997) and homology searching (Krogh *et al.*, 1994b; Eddy, 1998; Karplus *et al.*, 1998), but have been applied to a wide variety of problems (Churchill, 1989; Eddy, 1995; Thorne *et al.*, 1996). While fundamentally more powerful models, such as stochastic context-free grammars, are preferable for certain applications (Durbin *et al.*, 1998), HMMs appear in many cases to strike the right balance between simplicity and expressiveness.

[1]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, CA 95064.
[2]Howard Hughes Medical Institute, University of California, Santa Cruz, CA 95064.

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 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 (Edwards and Cavalli-Sforza, 1964; Neyman, 1971; Felsenstein, 1973, 1981; Yang, 1993; Whelan *et al.*, 2001). 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 (1995), 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 (1996) and Yang (1995) 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 (Thorne *et al.*, 1996; Goldman *et al.*, 1996; Liò *et al.*, 1998) 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. More recently, Husmeier and Wright (2001) have applied combined phylogenetic HMMs to the problem of detecting recombination events. (Somewhat different combinations of HMMs and phylogeny have also been applied to the problems of multiple alignment [Holmes and Bruno, 2001] and combined tree-building/multiple alignment [Mitchison, 1999]). 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.* (1996) have pointed out, 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-dependent* models of base substitution, which consider the N bases preceding each base and are capable of capturing the dependence of substitution patterns on neighboring bases (Blake *et al.*, 1992; Morton *et al.*, 1997). 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 dependence) 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 $x_{i,j}$ belongs to an alphabet $\boldsymbol{\Sigma}$, e.g., $\boldsymbol{\Sigma}$ = (A,C,G,T) (it is convenient to impose an ordering on the alphabet). 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 $|\boldsymbol{\Sigma}| \times |\boldsymbol{\Sigma}|$ and defines a continuous-time Markov process for base substitution (to be described 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 is unrooted, and effectively has one fewer node and edge. The vector $\boldsymbol{\beta}$ assigns a nonnegative real value to each branch of the tree, representing its evolutionary length, usually as an expected number of substitutions per site (see below). The vector $\boldsymbol{\pi}$ is of dimension $|\boldsymbol{\Sigma}|$ 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.

2.1. Computing the likelihood of a tree model

The critical step of computing the likelihood of a given tree model, $P(\mathbf{X}|\boldsymbol{\psi})$ is accomplished using the "pruning" algorithm of Felsenstein (1973, 1981). The sites of the alignment are assumed independent, so that $P(\mathbf{X}|\boldsymbol{\psi}) = \prod_{i=1}^{L} P(\mathbf{X}_i|\boldsymbol{\psi})$. The probability of each site is $P(\mathbf{X}_i|\boldsymbol{\psi}) = \sum_{\mathcal{L}} P(\mathcal{L}, \mathbf{X}_i|\boldsymbol{\psi})$, where $\mathcal{L}$ is a labeling of the $n - 1$ ancestral nodes of the tree with elements from $\boldsymbol{\Sigma}$ (the labels at the leaves are fixed by $\mathbf{X}_i$). Felsenstein's algorithm uses dynamic programming to compute this sum efficiently, as follows. Let u be any node in $\boldsymbol{\tau}$, and let v and w be its children. In addition, let t_v and t_w be the lengths of the branches connecting u to v and u to w, respectively. Suppose that we can compute the probability of base b replacing base a over a branch of length t, which we will denote $P(b|a, t)$ (see below). Now, following Durbin *et al.*, 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 is defined by the recursion

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

where I is the indicator function and x_u is the element of $\mathbf{X}_i$ corresponding to leaf u. The total probability of $\mathbf{X}_i$ is given by $P(\mathbf{X}_i|\boldsymbol{\psi}) = \sum_a \pi_a P(L_r|a)$, where r is the root of the tree. Felsenstein's algorithm allows the likelihood of the model to be computed in time linear in the size of the alignment (assuming $|\boldsymbol{\Sigma}|$ is a small constant).

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 an EM or generic nonlinear optimization algorithm (e.g., a quasi-Newton or conjugate gradients 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 $a, b \in \boldsymbol{\Sigma}$ and any nonnegative 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 $\mathbf{Q}$ (Yang *et al.*, 1994; Whelan *et al.*, 2001). 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, $\mathbf{Q} = \{q_{i,j}\}$ $(1 \leq i, j \leq |\mathbf{\Sigma}|)$ 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 $|\mathbf{\Sigma}| = 4$ (DNA), $\mathbf{Q}$ 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 in wide use is that of Hasegawa, Kishino, and Yano (1985) (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 $q_{i,i} = -\sum_{j:j\neq i} q_{i,j}$.

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. This matrix is given by $\mathbf{P}(t) = e^{\mathbf{Q}t}$, where $e^{\mathbf{Q}t} = \sum_{i=0}^{\infty} \frac{(\mathbf{Q}t)^i}{i!}$ (Karlin and Taylor, 1975; Liò and Goldman, 1998). Matrix $\mathbf{Q}$ is generally diagonalizable as $\mathbf{Q} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}$, 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 substitutions at each site occur at the same rate. Variation in the rate of substitution can be allowed for by introducing an appropriately distributed random scaling factor for the branch lengths. Yang has developed a method in which a gamma distribution is assumed and its shape parameter α is estimated from the data (Yang, 1993, 1994). For computational tractability, the gamma distribution is discretized into k rate categories, and a rate constant r_j $(1 \leq j \leq k)$ is computed for each category (the values $r_1, \ldots, r_k$ are a function of α). The probability $P(\mathbf{X}_i|\boldsymbol{\psi})$ can be approximated as

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

which can be computed with k invocations of Felsenstein's algorithm. Yang showed that the improvement in likelihoods drops off quickly after about $k = 3$ and recommended $k = 4$ as an appropriate choice.

Felsenstein and Churchill (1996) observed 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, and they introduced a model that also uses a discrete set of rate categories, but additionally assumes sites are assigned to categories by a Markov process. This process is defined by an autocorrelation parameter λ:

[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 $\sum_{i,j:i\neq j} \pi_i q_{i,j} = 1$.

if column $\mathbf{X}_{i-1}$ is assigned to category j, then with probability λ, column $\mathbf{X}_i$ will be assigned to category j, and with probability $1-\lambda$, $\mathbf{X}_i$ will be assigned to 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 (Durbin *et al.*, 1998). Similarly, the maximum-likelihood assignment of sites to categories can be obtained by the Viterbi algorithm, and the posterior probability that each site is assigned to 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 probabilities between them according to a bivariate gamma distribution. Thus, the parameter α (a free parameter in the fitting procedure) defines the rate categories (Felsenstein and Churchill left them to be set by the user). 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 [Press *et al.*, 1992]). 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 [1995]). Our model reduces to the discrete gamma model when $\lambda = 0$.

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 know the probability that any site $\mathbf{X}_i$ obeys model $\boldsymbol{\psi}_j$, given that the previous site, $\mathbf{X}_{i-1}$, obeys model $\boldsymbol{\psi}_{j'}$. 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 first codon positions, others of sites in second and third codon positions, and another of noncoding sites; the Markov transition probabilities might reflect the high probability that a second codon position follows a first codon position, the low but nonzero probability that a noncoding site follows a second codon position, and the zero probability that a third codon position follows a first codon position. Alternatively, the functional categories might correspond to secondary structural characteristics, as in the models of Thorne, Goldman, *et al.* (Thorne *et al.*, 1996; Goldman *et al.*, 1996; Liò *et al.*, 1998), or to any biological property that changes in some nonrandom 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 can be useful in detecting recombination (Husmeier and Wright, 2001).

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

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

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. 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) (Durbin *et al.*, 1998).

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 (Yang, 1996). 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 overall 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 (Durbin *et al.*, 1998). 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 reestimated on each iteration of the algorithm.

Having presented a generalization of rate categories, 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 (Matassi *et al.*, 1999; Williams and Hurst, 2000; Mouse Genome Sequencing Consortium, 2002). Rate categories, therefore, are in a sense orthogonal to functional categories, and it may be useful when inferring functional categories to allow changes in rate within each functional class (for example, to allow a slow-evolving noncoding region to be distinguished 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 $1 \leq i' \leq k$,

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

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 $1 \leq i', j' \leq k$,

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

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 converted to a representation 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 arise 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 (Mitchison, 1999). 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 by Felsenstein (1973) and has been implemented in PHYLIP (Felsenstein, 1993) and PAML (Yang, 1997), 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 $\mathbf{X}_i = (\mathrm{A,C,C,*,A})^{\mathrm{T}}$ has $M = \{(\mathrm{A,C,C,A,A})^{\mathrm{T}}, (\mathrm{A,C,C,C,A})^{\mathrm{T}}, (\mathrm{A,C,C,G,A})^{\mathrm{T}}, (\mathrm{A,C,C,T,A})^{\mathrm{T}}\}$. 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

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

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 [Krogh, 1997; Burge and Karlin, 1997]). 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-dependent 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 $|\mathbf{\Sigma}|^{N+1}$, consisting of all $(N+1)$-tuples of characters in $\mathbf{\Sigma}$, and adjust the dimensions of $\mathbf{Q}$ (the rate matrix) and $\boldsymbol{\pi}$ (the vector of equilibrium frequencies) accordingly. This is essentially what has been done in codon-based models (Goldman and Yang, 1994; Muse and Gaut, 1994; Pedersen *et al.*, 1998) and models designed to accommodate paired bases in RNA genes (Schöniger and von Haeseler, 1994; Muse, 1995; Rzhetsky, 1995; Tillier and Collins, 1995). In the case of $N = 1$, corresponding to dinucleotides, Felsenstein's algorithm can be written

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

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. 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 in 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 $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_q$ has been estimated from labeled training data. To apply the model to an unlabeled dataset, 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. 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 be computed only 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

[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 $\mathbf{Z} = (*, \ldots, *)^{\mathrm{T}}$.

for such changes [Averof *et al.*, 2000]; as future work, we intend to relax this restriction). Thus, R2 is defined with $\mathbf{Q} = \{q_{i,j}\}$ such that

$$q_{i,j} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ by } > 1 \text{ nucleotide} \\ a_{i,j}\pi_j & \text{if } i \text{ and } j \text{ differ by 1 nucleotide} \\ -\sum_{k:k\neq i} q_{i,k} & \text{if } i = j \end{cases}$$

where the variables of the form $a_{i,j}$ represent free parameters and $a_{i,j} = a_{j,i}$ for all $1 \leq i, j \leq |\mathbf{\Sigma}|^2$ (there are a total of 48 free parameters). Matrix R2S is identical to R2, except for the constraint that $a_{i,j} = a_{i',j'}$ if i' is the reverse complement of i and j' is the reverse complement of j, which cuts the number of parameters in half to 24. Matrix 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 the BFGS algorithm (Press *et al.*, 1992), with gradients computed using the difference method. The code was shown to give equivalent results to those of the PAML package (Yang, 1997) 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 implemented. The software is available upon request.

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 eight other eutherian mammals (Thomas *et al.*, 2003) (see species in Fig. 1). The sequences are products of the NISC Comparative Sequencing Program (www.nisc.nih.gov). The multiple alignment, which is derived from pairwise local alignments, was produced with the MultiPipMaker program (Schwartz *et al.*, 2003). It contains 2.3 million columns,

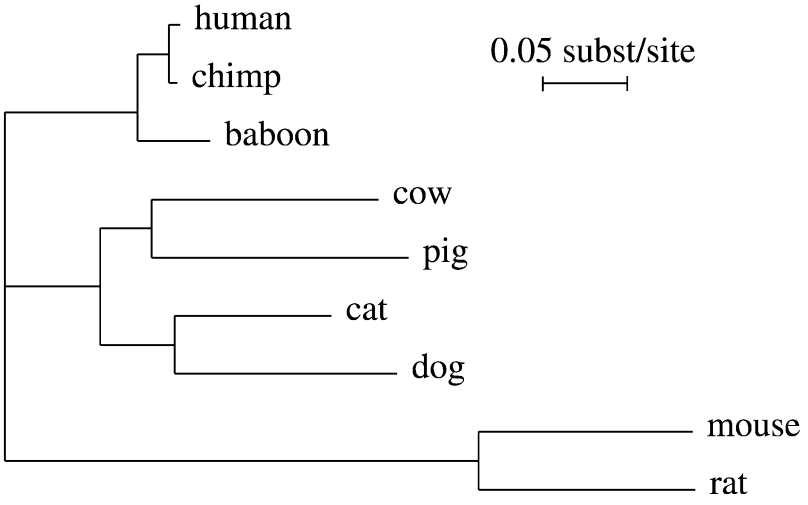

FIG. 1. Phylogenetic tree assumed for the nine species (unrooted). Branch lengths are drawn in the proportions estimated for the AR alignment using the REV model (with discrete gamma).

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), 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 (Thomas *et al.*, 2003). The annotations include ten 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 (*www.ftp.genome.washington.edu/cgi-bin/RepeatMasker*) 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 (2002). More such sites were available than we could efficiently analyze with our current software, so we selected 20,000 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. Of these, 1,083 (2.2%) 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.

3. RESULTS

The phylogenetic tree relating the nine species was assumed to have the (unrooted) topology shown in Fig. 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 this 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 (Murphy *et al.*, 2001; Thomas *et al.*, 2003).

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 (Huelsenbeck and Rannala, 1997). 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) exceeds all others by an order of magnitude. The LRT is applicable only 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.[3] 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. This improvement holds up well in cross-validation experiments and in comparisons based on the Bayesian information criterion (Siepel and Haussler, 2004a).

[3]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.

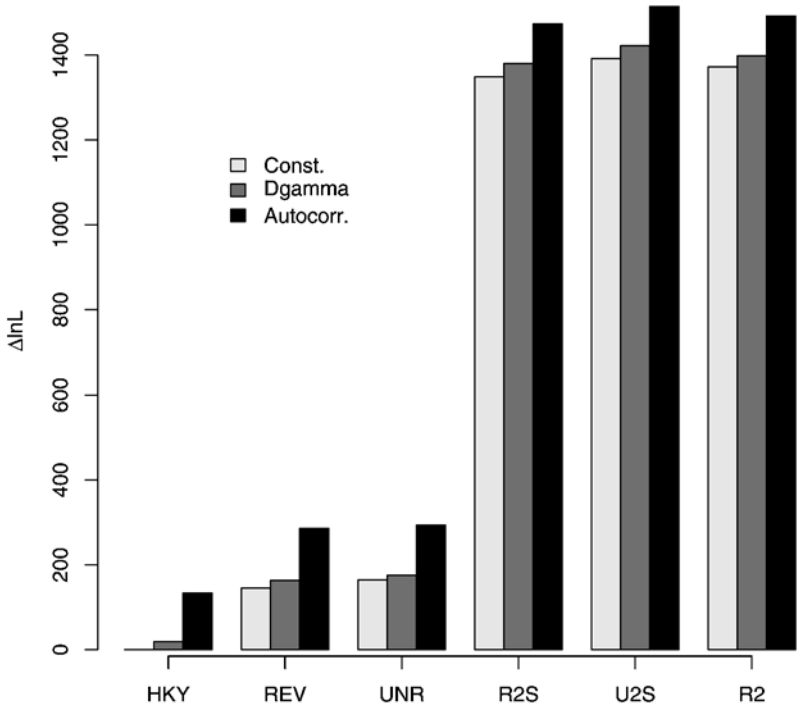

FIG. 2. Log likelihoods for the AR alignment under various substitution and rate models. Values are shown relative to HKY with constant rates.

Estimates of key parameters were generally similar under all models. The branch lengths were highly consistent, although they did tend to increase slightly with models of increasing richness, as has been noted previously (Yang *et al.*, 1994). 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 (5.7 [HKY], 6.0 [REV], and 6.1 [UNR]), but increased significantly when switching to dinucleotide substitution models (8.6 [R2S], 8.7 [U2S], and 9.3 [R2]).[4] This may be partly because some apparent rate variation can be explained by considering the context of each substitution, but could also be an artifact of applying the discrete gamma model to pairs of sites rather than individual sites. 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 [Lewin, 2000]); 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. The estimated rates of CpG transversions (CG→AG and CG→GG) were also considerably higher than those of other transversions. A variety of other, more subtle, differences in estimated rates could be seen, suggesting that the pattern of context-dependent substitution is complex (see Siepel and Haussler [2004a] for a more detailed discussion of these issues).

To contrast coding and neutral DNA, we performed a similar experiment with all bases in the second codon position of the entire two-megabase alignment (Fig. 3) (we did not apply the autocorrelation model to this dataset, 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 Figs. 2 and 3). The estimated rate matrix (not shown) appears to capture something about the pattern of amino acid substitutions, but the effect is

[4]Estimates of α were quite high in all cases, consistent with the assumption of neutral evolution. The low rate variation in this dataset explains the relatively small improvement seen with the discrete gamma model.

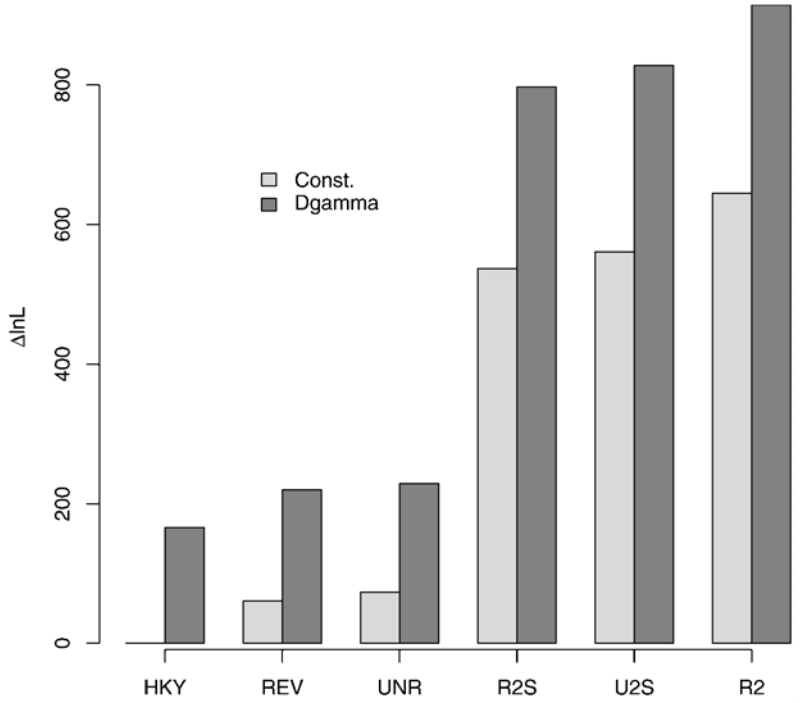

FIG. 3. Log likelihoods for sites in the second codon position, under various substitution and rate models. Values are shown relative to HKY with constant rates.

imperfect, because many dinucleotide substitutions correspond to mixtures of quite different amino acid substitutions (see Discussion). 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 first-, second-, or third-codon positions, or "other," and for the the six-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 changes in labels; no pseudocounts were used. 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 when moving both from one-category to four-category models and from four-category to six-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 were all greater than 0.99).[5] 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).

[5]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 one-state models, strong autocorrelation in noncoding regions overwhelms weak autocorrelation in coding regions (recall that the latter make up only about 2% of all sites).

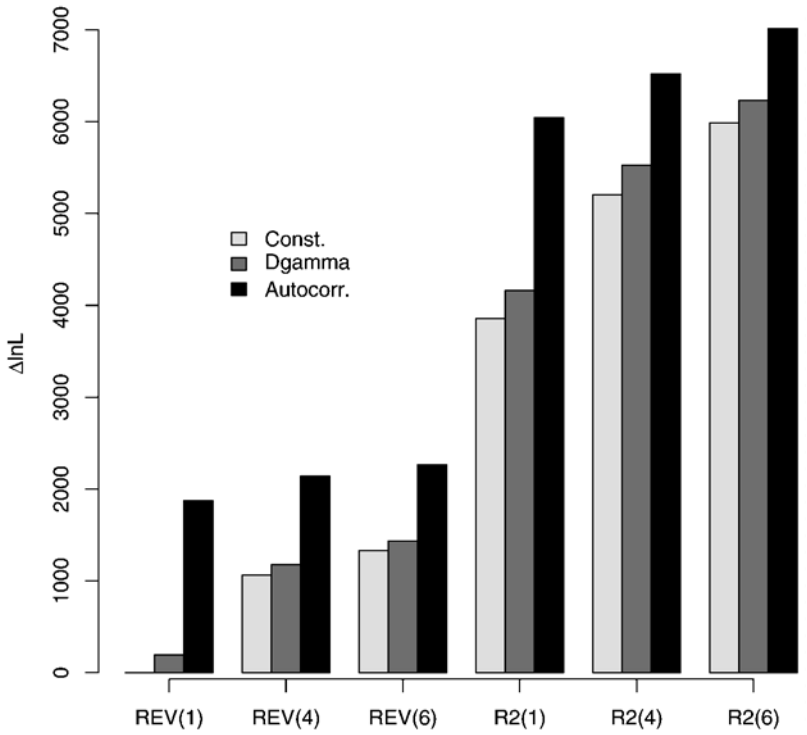

FIG. 4. Log likelihoods for the WNT2 alignment, for the REV and R2 substitution models, three models of rate variation, and sets of one, four, and six functional categories. Values are shown relative to REV with constant rates and one functional category.

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-dependent 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 effective. Judging by the case of dinucleotides, context-dependent 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 tuple size of three. The methods described here can be extended to nucleotide triples, but at the cost of much larger numbers of free parameters and more expensive likelihood computations, demanding alternative strategies for parameter estimation. In recent work, we have shown that considering nucleotide triples, rather than dinucleotides, can produce a substantial further improvement in goodness of fit, for both coding and noncoding data (Siepel and Haussler, 2004a).

Our approach to the problem of context dependence is limited in certain ways compared to strategies that have recently been proposed for the two-sequence case (Jensen and Pedersen, 2000; Pedersen and Jensen, 2001; Arndt *et al.*, 2002). For example, our model imposes unidirectional dependence of sites rather than allowing dependencies to flow in both directions, and it does not permit context effects to "cascade" beyond the limits of an $(N + 1)$-tuple along each branch of the tree. In addition, our model is not faithful to the assumed process in that it does not require the ancestral states associated with overlapping tuples to be consistent. Nevertheless, it is a valid probabilistic model in the sense that the probabilities of all possible inputs (of a given size) sum to one, and it has the advantage over "process-based" models of allowing likelihoods to be computed exactly and efficiently. As future work, we hope to compare this "empirical" model with process-based models, using MCMC or one of the approximation techniques available for graphical models with "loops" of dependency (Murphy *et al.*, 1999; Wainwright *et al.*, 2001; Yedidia *et al.*, 2001).

We have recently learned of independent work on combined phylogenetic and hidden Markov models by Pedersen and Hein (2003) and McAuliffe *et al.* (2003), specifically focused on gene prediction. Pedersen and Hein used a codon-based phylogenetic model for coding states and used models that assume independent sites for noncoding states; McAuliffe *et al.* used only independent-site models. These papers include some extensions not discussed here (e.g., McAuliffe *et al.* combine phylogenetic models with a *generalized* hidden Markov model, which allows for arbitrary length distributions of features such as exons and introns) and report encouraging gene-prediction performance in preliminary experiments. However, it appears that gene-prediction performance can be improved substantially by incorporating higher-order states into a combined phylogenetic HMM (Siepel and Haussler, 2004b).

It is worth noting that combined phylogenetic and hidden Markov models could potentially be useful in a wide variety of bioinformatics problems, including multiple alignment (as noted by Pedersen and Hein [2003]) and homology searching. We have focused in this paper on DNA sequences, but phylogenetic HMMs could be used as well with amino acid sequences, given appropriate models of substitution.

5. ACKNOWLEDGMENTS

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 dataset. 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.).

REFERENCES

Arndt, P.F., Burge, C.B., and Hwa, T. 2002. DNA sequence evolution with neighbor-dependent mutation. *Proc. 6th Int. Conf. Computational Biology*, 32–38.

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

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

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

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

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

Eddy, S. 1995. Multiple alignment using hidden Markov models. *Proc. 3rd Int. Conf. Intelligent Systems for Molecular Biology*, 114–120.

Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* 14, 755–763.

Edwards, A.W.F., and Cavalli-Sforza, L.L. 1964. Reconstruction of evolutionary trees, 67–76, *in* V.H. Heywood, and J. McNeill, eds., *Phenetic and Phylogenetic Classification*, Systematics Association, London.

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

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

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

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

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

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

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

Holmes, I., and Bruno, W.J. 2001. Evolutionary HMMs: A Bayesian approach to multiple alignment. *Bioinformatics* 17, 803–820.

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

Husmeier, D., and Wright, F. 2001. Detection of recombination in DNA multiple alignments with hidden Markov models. *J. Comp. Biol.* 8, 401–427.

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

Karlin, S., and Taylor, H.M. 1975. *A First Course in Stochastic Processes*, 2nd ed., Academic Press, NY.

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

Krogh, A. 1997. Two methods for improving performance of an HMM and their application for gene finding. *Proc. 5th Int. Conf. Intelligent Systems for Molecular Biology*, 179–186.

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

Krogh, A., Mian, I.S., and Haussler, D. 1994a. A hidden Markov model that finds genes in *E. coli* DNA. *Nucl. Acids Res.* 22, 4768–4778.

Kulp, D., Haussler, D., Reese, M.G., and Eeckman, F.H. 1996. A generalized hidden Markov model for the recognition of human genes in DNA. *Proc. 4th Int. Conf. Intelligent Systems for Molecular Biology*, 134–142.

Lewin, B. 2000. *Genes VII*, Oxford University Press, Oxford, UK.

Liò, P., and Goldman, N. 1998. Models of molecular evolution and phylogeny. *Genome Res.* 8, 1233–1244.

Liò, P., Goldman, N., Thorne, J.L., and Jones, D.T. 1998. PASSML: Combining evolutionary inference and protein secondary structure prediction. *Bioinformatics* 14, 726–733.

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

McAuliffe, J.D., Pachter, L., and Jordan, M.I. 2003. *Multiple-sequence functional annotation and the generalized hidden Markov phylogeny*, *in* Technical report 647, Department of Statistics, University of California, Berkeley.

Mitchison, G.J. 1999. A probabilistic treatment of phylogeny and sequence alignment. *J. Mol. Evol.* 49, 11–22.

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

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

Murphy, K., Weiss, Y., and Jordan, M.I. 1999. Loopy belief-propagation for approximate inference: An empirical study. *Proc. 15th Conf. Uncertainty in Artificial Intelligence (UAI)*.

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

Muse, S. 1995. Evolutionary analyses of DNA sequences subject to constraints on secondary structure. *Genetics* 139, 1429–1439.

Muse, S.V., and Gaut, B.S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the choroplast genome. *Mol. Biol. Evol.* 11, 715–724.

Neyman, J. 1971. Molecular studies of evolution: A source of novel statistical problems, *in* S.S. Gupta and J. Yackel, eds., *Statistical Decision Theory and Related Topics*, Academic Press, NY.

Pedersen, A.-M.K., and Jensen, J.L. 2001. 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.

Pedersen, A.-M.K., Wiuf, C., and Christiansen, F.B. 1998. A codon-based model designed to describe lentiviral evolution. *Mol. Biol. Evol.* 15, 1069–1081.

Pedersen, J.S., and Hein, J. 2003. Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics* 19, 219–227.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. 1992. *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed., Cambridge University Press, London.

Rzhetsky, A. 1995. Estimating substitution rates in ribosomal RNA genes. *Genetics* 141, 771–783.

Schöniger, M., and von Haeseler, A. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phyl. Evol.* 3, 240–247.

Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A., NISC Comparative Sequencing Program, Green, E.D., Hardison, R.C., and Miller, W. 2003. MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucl. Acids Res.* 31, 3518–3524.

Siepel, A., and Haussler, D. 2004a. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* 21, 468–488.

Siepel, A., and Haussler, D. 2004b. Computational identification of evolutionarily conserved exons. *Proc. 8th Int. Conf. Computational Molecular Biology*. To appear.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., *et al.* 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424, 788–793.

Thorne, J.L., Goldman, N., and Jones, D.T. 1996. Combining protein evolution and secondary structure. *Mol. Biol. Evol.* 13, 666–673.

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

Wainwright, M., Jaakkola, T., and Willsky, A. 2001. Tree-based reparameterization for approximate estimation on loopy graphs, *in Advances in Neural Information Processing Systems 14*, MIT Press, Cambridge, MA.

Whelan, S., Liò, P., and Goldman, N. 2001. Molecular phylogenetics: State-of-the-art methods for looking into the past. *Trends Genet.* 17, 262–272.

Williams, E.J.B., and Hurst, L.D. 2000. The proteins of linked genes evolve at similar rates. *Nature* 407, 900–903.

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* 139, 993–1005.

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

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

Yang, Z., Goldman, N., and Friday, A. 1994. Comparison of models for nucleotide substution used in maximum likelihood phylogenetic estimation. *Mol. Biol. Evol.* 11, 316–324.

Yedidia, J., Freeman, W., and Weiss, Y. 2001. *Bethe free energy, Kikuchi approximations, and belief propagation algorithms*, *in* Technical report TR2001-16, Mitsubishi Electronic Research Laboratories.

Address correspondence to:
Adam Siepel
Center for Biomolecular Science and Engineering
Baskin Engineering Building
University of California
1156 High Street
Santa Cruz, CA 95064

E-mail: acs@soe.ucsc.edu