

Phylogenetic Estimation of Context-Dependent Substitution Rates by Maximum Likelihood

Adam Siepel* and David Haussler*†

*Center for Biomolecular Science and Engineering, University of California, Santa Cruz; †Howard Hughes Medical Institute, University of California, Santa Cruz

Nucleotide substitution in both coding and noncoding regions is context-dependent, in the sense that substitution rates depend on the identity of neighboring bases. Context-dependent substitution has been modeled in the case of two sequences and an unrooted phylogenetic tree, but it has only been accommodated in limited ways with more general phylogenies. In this article, extensions are presented to standard phylogenetic models that allow for better handling of context-dependent substitution, yet still permit exact inference at reasonable computational cost. The new models improve goodness of fit substantially for both coding and noncoding data. Considering context dependence leads to much larger improvements than does using a richer substitution model or allowing for rate variation across sites, under the assumption of site independence. The observed improvements appear to derive from three separate properties of the models: their explicit characterization of context-dependent substitution within N -tuples of adjacent sites, their ability to accommodate overlapping N -tuples, and their rich parameterization of the substitution process. Parameter estimation is accomplished using an expectation maximization algorithm, with a quasi-Newton algorithm for the maximization step; this approach is shown to be preferable to ordinary Newton methods for parameter-rich models. Overlapping tuples are efficiently handled by assuming Markov dependence of the observed bases at each site on those at the $N - 1$ preceding sites, and the required conditional probabilities are computed with an extension of Felsenstein's algorithm. Estimated substitution rates based on a data set of about 160,000 noncoding sites in mammalian genomes indicate a pronounced CpG effect, but they also suggest a complex overall pattern of context-dependent substitution, comprising a variety of subtle effects. Estimates based on about 3 million sites in coding regions demonstrate that amino acid substitution rates can be learned at the nucleotide level, and suggest that context effects across codon boundaries are significant.

Introduction

Many of the simplifying assumptions originally introduced for phylogenetic inference have been relaxed, resulting in models of improved realism and power. For example, the assumptions that each nucleotide or amino acid substitutes for the other at the same average rate, and that this average rate is constant across sites (Neyman 1971; Felsenstein 1981), have both given way to more realistic alternatives (Whelan, Liò, and Goldman 2001). One assumption that has largely been maintained, however—despite general agreement that it is biologically unrealistic—is that the sites at which evolution occurs can be considered independent.

Various methods have been proposed for relaxing site independence in limited ways, without sacrificing the essential strengths of maximum-likelihood methods for phylogenetic inference. For example, models have been introduced that allow for dependence between sites of the same codon (Goldman and Yang 1994; Muse and Gaut 1994; Pedersen, Wiuf, and Christiansen 1998; Yang et al. 2000; Schadt and Lange 2002), correlation of the overall rates of substitution at adjacent sites (Felsenstein and Churchill 1996; Yang 1995), and dependence between paired bases in ribosomal RNA genes (Schöniger and von Haeseler 1994; Muse 1995; Rzhetsky 1995; Tillier and Collins 1995). In addition, models have been developed for the special case of two sequences and a reversible substitution process that allow for general context-de-

pendent substitution, with substitution rates for each base depending on the identity of flanking bases (Jensen and Pedersen 2000; Pedersen and Jensen 2001). With these models, the likelihood computation can no longer be expressed as a product over the sites of an alignment, and exact parameter estimation becomes intractable (a Markov random field arises from the bidirectional dependencies between adjacent sites); Markov chain Monte Carlo (MCMC) or similar methods must be used instead. (Another, simpler model of context-dependent substitution for pairs of sequences has been developed by Arndt, Burge and Hwa (2002), along with an approximation procedure, adapted from nonlinear dynamics, that allows for efficient parameter estimation.) These models accurately reflect an assumed *process* of context-dependent substitution, and therefore, will be described here as “process-based,” in contrast to more empirical models (see below). To our knowledge, process-based models for context-dependent substitution have not yet been extended for use with a general phylogenetic tree.

It is now widely agreed that site independence is an unacceptable assumption in coding regions, where selection acts primarily at the level of amino acids, which are coded for by triplets of adjacent nucleotides. Indeed, the inadequacy of this assumption is the premise underlying codon models, which do appear to improve significantly on the goodness of fit of independent-site models in coding regions (Goldman and Yang 1994). Substitution rates in noncoding DNA, however, are also highly dependent on neighboring bases, and this phenomenon has mostly been ignored in phylogenetic analysis (except that sites subject to the strongest dependencies are often discarded).

Studies of aligned gene/pseudogene pairs in primates have revealed strong evidence of context-dependent

Key words: neighbor-dependent substitution, CpG effect, codon model, expectation maximization, substitution rate matrix.

E-mail: acs@soe.ucsc.edu.

Mol. Biol. Evol. 21(3):468–488, 2004

DOI: 10.1093/molbev/msh039

Advance Access publication December 5, 2003

Molecular Biology and Evolution vol. 21 no. 3

© Society for Molecular Biology and Evolution 2004; all rights reserved.

substitution in (presumably) neutrally evolving regions, with the rate at which one nucleotide substitutes for another varying as much as 10-fold for different sets of flanking bases, and context-dependent rates overall spanning a more than 50-fold range (Blake, Hess, and Nicholson-Tuell 1992; Hess, Blake, and Blake 1994). Similar effects have been observed recently in connection with single-nucleotide polymorphisms (SNPs) in the human genome (Zhao and Boerwinkle 2002). The strongest context effects in these studies have been seen with C→T transitions in CpG dinucleotides (or G→A transitions on the reverse strand), and are presumably due to methylation and spontaneous deamination of cytosines (Bulmer 1986; Ehrlich, Zhang, and Inamdar 1990). Other effects have been noted as well, however, such as a tendency for higher rates of substitution where purines and pyrimidines alternate, an increased rate of T→C transitions in TpA dinucleotides, and a higher than expected rate of C→A and C→G (G→T and G→A) transversions in CpG dinucleotides (Blake, Hess, and Nicholson-Tuell 1992; Hess, Blake, and Blake 1994). A somewhat weak dependence of the transition/transversion ratio on the A + T content of flanking bases has also been observed in primates (Zhao and Boerwinkle 2002; Yang, Chen, and Li 2002), echoing the stronger dependence observed in the chloroplast and mitochondrial genomes of plants (Morton and Clegg 1995; Morton, Oberholzer, and Clegg 1997; Yang, Chen, and Li 2002). Context effects are strongest for the two immediate neighbors of a given site, but under certain conditions, they can be detected as well for the second and third pairs of flanking bases (Morton 1997; Morton, Oberholzer, and Clegg 1997). Subtle effects may extend as far as 200 bp (Zhao and Boerwinkle 2002), and longer-range dependencies of substitution rates on base composition and other regional properties have also been observed (Bernardi 2000; Hardison, Roskin, and Yang 2003). Context dependence of substitution rates is thought to be the result of neighbor effects on various phenomena, including polymerase fidelity and proofreading, mismatch repair, and mismatch stability (briefly reviewed by Blake, Hess, and Nicholson-Tuell 1992; Morton, Oberholzer, and Clegg 1997).

Most studies of context-dependent substitution rates in noncoding DNA have used simple counting methods and pairwise alignments of highly similar sequences. These methods have some clear deficiencies: for example, they do not allow for multiple substitutions per site or differences in evolutionary distance between aligned pairs of sequences (e.g., in different gene/pseudogene pairs), and they do not benefit from consideration of multiple homologous sequences and a phylogeny. If context-dependent substitution could be incorporated into probabilistic phylogenetic models, better estimates of context-dependent substitution rates might be obtained, as well as improved estimates of branch lengths and (possibly) tree topologies.

In this article, methods are introduced for incorporating context-dependent substitution into phylogenetic models that are more closely related to methods used in computational gene-finding (Durbin et al. 1998; Siepel and Haussler 2004) than to those used with process-based models. The basic strategy here is to extend and generalize

codon models for improved handling of context dependence, without sacrificing too much in the way of computational efficiency. We stop short of a process-based description of context-dependent substitution, and as a result are able, for the most part, to retain the standard framework for likelihood computation and parameter estimation (the need for sampling or approximation algorithms is avoided). We begin with the introduction of arbitrary “ N -th order” models, defined in terms of N -tuples of characters (for small N), and then a discussion of how these models can be fitted to large data sets using an expectation maximization (EM) algorithm. Next, a method is introduced, based on an assumption of Markov dependence of each site on its $N - 1$ predecessors, that allows N -tuples to overlap, so that context effects can be captured between all adjacent pairs of sites. Results are then presented for two large data sets of mammalian sequence—one of about 160,000 sites in noncoding regions and one of about 3 million sites in coding regions—which demonstrate that our context-dependent models produce significant improvements in goodness of fit compared with both codon models and independent-site models. Allowing for overlapping tuples of sites is shown to be important, as is using a sufficiently rich parameterization of the substitution process. Parameter estimates are presented for context-dependent substitution in both noncoding and coding regions, and discussed in some detail.

Materials and Methods

Background and Notation

Let a *phylogenetic model* $\psi = (\mathbf{Q}, \boldsymbol{\pi}, \tau, \boldsymbol{\beta})$ be a four-tuple consisting of a substitution rate matrix \mathbf{Q} , a vector of equilibrium frequencies $\boldsymbol{\pi}$, a (rooted) binary tree $\tau = (\mathcal{V}, \mathcal{E})$ (where \mathcal{V} and \mathcal{E} represent the nodes [vertices] and branches [edges] of the tree, respectively), and a set of branch lengths $\boldsymbol{\beta} = \{\beta_u \mid u \in \mathcal{V} - \{r\}\}$ (where β_u corresponds to the branch between node u and its parent, denoted $\sigma(u)$, and r is the root of the tree). The model is defined with respect to an alphabet Σ of size d , e.g., $\Sigma = (\text{A}, \text{C}, \text{G}, \text{T})$. In standard phylogenetic models, \mathbf{Q} has dimension $d \times d$, and $\boldsymbol{\pi}$ has dimension d . The tree has n leaves, or external nodes, corresponding to n present-day taxa; hence, in total there are $2n - 1$ nodes and $2n - 2$ branches. It is sometimes useful to distinguish between the set of leaves, $\mathcal{L} \subseteq \mathcal{V}$, and the set of internal nodes, $\mathcal{I} = \mathcal{V} - \mathcal{L}$. A rooted tree is assumed here for simplicity, but in practice an unrooted tree is often used; only minor changes to our methods are required for unrooted trees.

The parameters of a phylogenetic model are estimated with respect to a multiple alignment of n sequences, whose length (number of columns) we denote L . It is assumed that there has existed a sequence of length L , with characters drawn from Σ , corresponding to every node $u \in \mathcal{V}$; the alignment consists of the subset of these sequences that correspond to leaves of the tree. Let $\{X_{u,i}\}$ be a set of random variables, such that $X_{u,i}$ represents the i th character ($1 \leq i \leq L$) in the sequence corresponding to node $u \in \mathcal{V}$. The random variables corresponding to leaves of the tree are observed, with values given by the alignment, and the random variables corresponding to ancestral nodes are

unobserved, or latent. Let X_{\bullet} denote the set of observed variables, $X_{\bullet} = \{X_{u,i} | u \in \mathcal{L}\}$, and let X_{\circ} denote the set of latent variables, $X_{\circ} = \{X_{u,i} | u \in \mathcal{I}\}$. Let x_{\bullet} and x_{\circ} denote instances of X_{\bullet} and X_{\circ} , respectively, so that x_{\bullet} is completely defined by the given alignment, and x_{\circ} represents a set of possible ancestral sequences. Let $X_{\bullet,i}$ and $X_{\circ,i}$ be the sets of observed and latent variables, respectively, corresponding to column i in the alignment, $X_{\bullet,i} = \{X_{u,i} | u \in \mathcal{L}\}$ and $X_{\circ,i} = \{X_{u,i} | u \in \mathcal{I}\}$, and let $x_{\bullet,i}$ and $x_{\circ,i}$ be sets of instances of these variables.

The likelihood of a phylogenetic model ψ with respect to an alignment x_{\bullet} is obtained by summing over all possible values of the latent variables. Assuming site independence,

$$P(x_{\bullet} | \psi) = \prod_{i=1}^L P(x_{\bullet,i} | \psi) = \prod_{i=1}^L \sum_{x_{\circ,i}} P(x_{\bullet,i}, x_{\circ,i} | \psi).$$

The probability of an individual column, $P(x_{\bullet,i} | \psi) = \sum_{x_{\circ,i}} P(x_{\bullet,i}, x_{\circ,i} | \psi)$, can be computed with Felsenstein's "pruning" algorithm (Felsenstein 1981; Durbin et al. 1998) (see Appendix A), assuming that the probability is available of the substitution of any $b \in \Sigma$ for any $a \in \Sigma$ over a branch of length t ($t \geq 0$). These probabilities, denoted $P(b | a, t)$, are based on a continuous-time Markov model of substitution, defined by the rate matrix \mathbf{Q} . (The substitution process is assumed to be stationary and homogeneous.) Let the elements of \mathbf{Q} be denoted $\{q_{a,b}\}$ (for $a, b \in \Sigma$), with $q_{a,b}$ ($a \neq b$) representing the instantaneous rate at which a is replaced by b , and elements on the main diagonal defined such that each row sums to zero; i.e., $q_{a,a} = -\sum_{b \in \Sigma - \{a\}} q_{a,b}$. The probabilities $P(b | a, t)$ are given by the elements of the matrix $\mathbf{P}(t) = \exp(\mathbf{Q}t)$, where $\exp(\mathbf{Q}t) = \sum_{i=0}^{\infty} \frac{(\mathbf{Q}t)^i}{i!}$ (Karlin and Taylor 1975; Liò and Goldman 1998). (We will sometimes denote $P(b | a, t)$ as $[\exp(\mathbf{Q}t)]_{a,b}$, to be clear that it is a function of \mathbf{Q} .) \mathbf{Q} can be parameterized in various ways (Yang 1994a; Whelan, Liò, and Goldman 2001); in this article, we consider three of the standard parameterizations, corresponding to the HKY, REV (general reversible), and UNR (unrestricted) models (see table 1). By convention, \mathbf{Q} is scaled such that t can be interpreted as the expected number of substitutions per site (the scaling constraint is $\sum_{a,b \in \Sigma: a \neq b} \pi_a q_{a,b} = 1$).

A phylogenetic model can be extended easily to describe N -tuples of sites, for small N : the alphabet Σ is simply replaced with Σ^N , and the dimensions of \mathbf{Q} and π are adjusted accordingly. As long as the N -tuples can be assumed independent, the procedures for computing the likelihood of a model and estimating its parameters remain essentially unchanged, although they become more computationally intensive. (The matrix $\mathbf{P}(t) = \exp(\mathbf{Q}t)$ now describes the joint probabilities of substitution of N -tuples of bases.) We refer to such a model as having *order* N , and when $N > 1$, we call the model *context-dependent*. In this article, four kinds of N th order models are considered, corresponding to four parameterizations of the rate matrix \mathbf{Q} : an unrestricted model (UN), a reversible model (RN), a strand-symmetric unrestricted model (UNS), and a strand-symmetric reversible model (RNS) (see table 1). In all cases, the number of parameters is kept

Table 1
Summary of Nucleotide Substitution Models Discussed in This Article

Model	N^a	Parameters ^b	Description
HKY	1	1 + 3	Reversible single-nucleotide model of Hasegawa, Kishino, and Yano (1985), which allows only for non-uniform equilibrium frequencies and a transition/transversion bias
REV	1	6 + 3	General reversible single-nucleotide model (Tavaré 1986)
UNR	1	12	General unrestricted single-nucleotide model (Yang 1994a)
R2S	2	24 + 15	General reversible dinucleotide model with strand symmetry
R2	2	48 + 15	General reversible dinucleotide model
U2S	2	48	General unrestricted dinucleotide model with strand symmetry
U2	2	96	General unrestricted dinucleotide model
R3S	3	148 + 63	General reversible trinucleotide model with strand symmetry
R3	3	288 + 63	General reversible trinucleotide model
U3S	3	288	General unrestricted trinucleotide model with strand symmetry
U3	3	576	General unrestricted trinucleotide model
GYE	3	2 + 60	The codon model of Goldman and Yang (1994), with equal distances between amino acids
GYM	3	3 + 60	The codon model of Goldman and Yang (1994), with the physicochemical distance matrix of Miyata, Miyazawa, and Yasunaga (1979)

^a Order of the model.

^b Number of rate-matrix parameters + number of (free) equilibrium-frequency parameters (the latter are not required for unrestricted models). The scaling constraint on the rate matrix is ignored here. Note that these numbers do not characterize entire phylogenetic models (e.g., branch lengths are not considered). In context-dependent models ($N \geq 2$), instantaneous substitutions of more than one nucleotide are prohibited.

manageable by assuming an instantaneous rate of zero for substitutions of multiple nucleotides, as in most codon models. A scaling constraint of $\sum_{a,b \in \Sigma^N: a \neq b} \pi_a q_{a,b} = N$ allows for branch-length units of substitutions per site. See Appendix B for additional details.

For context-dependent models, the letter Z will be used in place of X to denote both N -tuples of sites and N -tuples of random variables. N -tuples will be assumed to be composed of adjacent sites, so that $Z_{u,j} = (X_{u,N(j-1)+1}, \dots, X_{u,Nj})$, $Z_{\bullet,j} = (X_{\bullet,N(j-1)+1}, \dots, X_{\bullet,Nj})$, and $Z_{\circ,j} = (X_{\circ,N(j-1)+1}, \dots, X_{\circ,Nj})$ (where $1 \leq j \leq L'$, $L' = L/N$). As with $x_{\bullet,i}$ and $x_{\circ,i}$, $z_{\bullet,j}$ and $z_{\circ,j}$ are instances of $Z_{\bullet,j}$ and

$Z_{\circ,j}$, respectively. To complete the parallel, the variables Z_{\bullet} , Z_{\circ} , z_{\bullet} , and z_{\circ} are defined, but are equivalent to X_{\bullet} , X_{\circ} , x_{\bullet} , and x_{\circ} , respectively. When N -tuples are independent (and nonoverlapping, as above), the likelihood of the model is simply $P(z_{\bullet} | \Psi) = \prod_{j=1}^L P(z_{\bullet,j} | \Psi)$. The probability of each tuple of columns, $P(z_{\bullet,j} | \Psi) = \sum_{z_{\circ,j}} P(z_{\bullet,j}, z_{\circ,j} | \Psi)$, can be computed with Felsenstein's algorithm, as before, but with substitution probabilities based on N -tuples of nucleotides. Overlapping N -tuples are discussed below.

Variation in the rate of evolution between different sites can be accommodated, with context-dependent as well as independent-site models, using the *discrete gamma* method of Yang (1993; 1994b). This method is based on the introduction of a random scaling parameter for the branch lengths β , which is assumed to have a gamma distribution, whose shape parameter α is estimated from the data. The distribution is partitioned into k "rate categories" of equal probability, each with a rate constant r_l ($1 \leq l \leq k$), and the probability $P(z_{\bullet,j} | \Psi)$ is approximated as

$$P(z_{\bullet,j} | \Psi) \approx \sum_{l=1}^k \frac{1}{k} \cdot P(z_{\bullet,j} | \mathbf{Q}, \boldsymbol{\pi}, \tau, r_l \boldsymbol{\beta}), \quad (1)$$

a quantity that can be computed with k invocations of Felsenstein's algorithm. Even quite small values of k appear to provide adequate approximations of the full continuous distribution; $k=4$ was recommended by Yang (1994b).

Estimating the Parameters of a Context-Dependent Model

A maximum-likelihood estimate (m.l.e.) of the parameters of a phylogenetic model, $\hat{\Psi} = \operatorname{argmax}_{\Psi} P(z_{\bullet} | \Psi)$, can be obtained by searching over tree topologies, and numerically optimizing $(\mathbf{Q}, \boldsymbol{\pi}, \boldsymbol{\beta})$ conditional on each possible τ . In this article, τ is assumed to be known, and the problem is simplified considerably. In addition, $\boldsymbol{\pi}$ is estimated directly from z_{\bullet} , as is the standard practice (each frequency π_a is estimated as the observed frequency of tuple a in z_{\bullet}), so that the critical problem is to obtain m.l.e.'s of \mathbf{Q} and $\boldsymbol{\beta}$ conditional on $\boldsymbol{\pi}$ and τ .

Many software packages, including PAML (Yang 1997), MOLPHY (Adachi and Hasegawa 1996), fastDNAML (Olsen et al. 1994), PHYLIP (Felsenstein 1993), and PAUP (Swofford 2002), use Newton-Raphson or quasi-Newton algorithms for parameter estimation (Press et al. 1992), computing partial derivatives of the likelihood function numerically or with a combination of numerical and analytical methods. Newton-based optimization algorithms are not well suited for models with richly parameterized rate matrices, however, because of the computational cost of computing partial derivatives of the likelihood function with respect to rate-matrix parameters (the same is true of any method that makes use of function derivatives, such as conjugate-gradient methods). Moreover, they tend to require large numbers of evaluations of the likelihood function, whose computational cost increases exponentially with the order of a context-dependent model, and linearly with the number of sites considered. We will show in this section that an EM algorithm (Dempster, Laird, and Rubin 1977) is generally preferable to quasi-Newton methods for context-dependent models, although quasi-Newton methods can be useful for

solving the optimization problem that arises on each iteration of the EM algorithm.

Suppose we seek the m.l.e. of $(\mathbf{Q}, \boldsymbol{\beta})$ given τ and $\boldsymbol{\pi}$: $(\hat{\mathbf{Q}}, \hat{\boldsymbol{\beta}}) = \operatorname{argmax}_{\mathbf{Q}, \boldsymbol{\beta}} P(z_{\bullet} | \mathbf{Q}, \boldsymbol{\pi}, \tau, \boldsymbol{\beta})$. The theorem behind the EM algorithm says that a (local) maximum of $P(z_{\bullet} | \mathbf{Q}, \boldsymbol{\pi}, \tau, \boldsymbol{\beta})$ can be obtained by iterative maximization of the expected value of $\log P(z_{\bullet}, z_{\circ} | \mathbf{Q}, \boldsymbol{\pi}, \tau, \boldsymbol{\beta})$, computed with respect to the (posterior) distribution of Z_{\circ} (Dempster, Laird, and Rubin 1977). It can be shown (Appendix C) that the general EM update rule, by which estimates for iteration $t+1$ are obtained from estimates for iteration t , here becomes

$$\begin{aligned} & (\hat{\mathbf{Q}}^{t+1}, \hat{\boldsymbol{\beta}}^{t+1}) \\ &= \operatorname{argmax}_{\mathbf{Q}, \boldsymbol{\beta}} \sum_{u \in \mathcal{V} - \{r\}} \sum_{a, b \in \Sigma^N} E[S(b, a, u) | z_{\bullet}, \hat{\mathbf{Q}}^t, \boldsymbol{\pi}, \tau, \hat{\boldsymbol{\beta}}^t] \\ & \quad \times \log([\exp(\mathbf{Q}\boldsymbol{\beta}_u)]_{a,b}) \end{aligned} \quad (2)$$

where $E[S(b, a, u) | z_{\bullet}, \hat{\mathbf{Q}}^t, \boldsymbol{\pi}, \tau, \hat{\boldsymbol{\beta}}^t]$ is the expected number of substitutions of $b \in \Sigma^N$ for $a \in \Sigma^N$ on the branch above u , given the observations at the leaves of the tree and the previously estimated parameters. Obtaining these values (the "E" step of the EM algorithm) can be accomplished with a procedure that is closely related to Felsenstein's algorithm (Appendix C). These expected values are then treated as constants when solving the remaining maximization problem (the "M" step).

This "inner" maximization problem remains non-trivial, but it can be solved numerically with the same kind of algorithm we have rejected for the larger problem of estimating $(\hat{\mathbf{Q}}, \hat{\boldsymbol{\beta}})$. Quasi-Newton methods (for example) are better suited for the inner problem than the outer one because the computational complexity of the function to be maximized is greatly reduced; in particular, the new function has no dependency on the length of the alignment (it requires $O(nd^{2N})$ time, versus $O(nL^l d^{2N})$ time for the complete likelihood function). The efficiency of the optimization procedure can be further improved by using analytical methods to obtain partial derivatives, approximating the derivatives with respect to rate-matrix parameters, and initializing the parameter values on each iteration with the solution of the previous iteration. The partial derivatives with respect to rate-matrix parameters can be derived using the Cauchy integral formula, as shown recently by Schadt and Lange (2002). See Appendix C for details.

Rate variation can also be accommodated with EM, using the discrete gamma model. In this case, the update rule becomes

$$\begin{aligned} & (\hat{\mathbf{Q}}^{t+1}, \hat{\boldsymbol{\beta}}^{t+1}, \hat{\boldsymbol{\alpha}}^{t+1}) \\ &= \operatorname{argmax}_{\mathbf{Q}, \boldsymbol{\beta}, \boldsymbol{\alpha}} \sum_{l=1}^k \sum_{u \in \mathcal{V} - \{r\}} \sum_{a, b \in \Sigma^N} E[S(b, a, u, l) | z_{\bullet}, \hat{\mathbf{Q}}^t, \boldsymbol{\pi}, \tau, \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\alpha}}^t] \\ & \quad \times \log([\exp(\mathbf{Q}r_l \boldsymbol{\beta}_u)]_{a,b}) \end{aligned} \quad (3)$$

where $E[S(b, a, u, l) | z_{\bullet}, \hat{\mathbf{Q}}^t, \boldsymbol{\pi}, \tau, \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\alpha}}^t]$ is the expected number of substitutions of $b \in \Sigma^N$ for base $a \in \Sigma^N$ on the branch above u for rate category l . This quantity reflects the posterior probabilities of the rate categories at each

site, as well as the posterior probabilities of substitutions. Note that the shape parameter α of the gamma distribution must be estimated as part of the inner maximization problem, and that the values r_1, \dots, r_k are implicitly functions of this parameter. With rate variation, the cost of the E step is increased by a factor of k , and the EM algorithm converges more slowly. See Appendix C for further details.

Markov Dependence Between Sites

It is possible to allow for overlapping tuples of columns, and hence to capture context effects between all adjacent columns of the alignment x_\bullet , by assuming Markov dependence of each column $x_{\bullet,i}$ on its $N - 1$ predecessors. In this case, the likelihood function can be written as

$$P(x_\bullet | \Psi) = P(x_{\bullet,1} | \Psi) \times P(x_{\bullet,2} | x_{\bullet,1}, \Psi) \cdots P(x_{\bullet,N-1} | x_{\bullet,1}, \dots, x_{\bullet,N-2}, \Psi) \times \prod_{i=N}^L P(x_{\bullet,i} | x_{\bullet,i-N+1}, \dots, x_{\bullet,i-1}, \Psi), \quad (4)$$

and the probability $P(x_{\bullet,i} | x_{\bullet,i-N+1}, \dots, x_{\bullet,i-1}, \Psi)$ can be computed as

$$P(x_{\bullet,i} | x_{\bullet,i-N+1}, \dots, x_{\bullet,i-1}, \Psi) = \frac{P(x_{\bullet,i-N+1}, \dots, x_{\bullet,i-1}, x_{\bullet,i} | \Psi)}{\sum_{\tilde{x}_{\bullet,i}} P(x_{\bullet,i-N+1}, \dots, x_{\bullet,i-1}, \tilde{x}_{\bullet,i} | \Psi)}, \quad (5)$$

where $\sum_{\tilde{x}_{\bullet,i}}$ represents a sum over all possible columns of observed characters at site i (all possible values at the leaves of the tree). The joint probability in the numerator of equation 5 is simply the probability of an observed N -tuple of sites, and hence can be computed using an N th order phylogenetic model. The sum in the denominator can be computed with the same model, using a slight adaptation of Felsenstein's algorithm that sums over all possible ancestral states at sites $i - N + 1, \dots, i$, as usual, but also over all possible values of $x_{\bullet,i}$ (see details in Appendix B). The new algorithm uses the same principle that is used in the version of Felsenstein's algorithm that allows for missing data (Appendix A), and as a result, it differs from Felsenstein's algorithm only in its initialization step. Thus, the conditional probability $P(x_{\bullet,i} | x_{\bullet,i-N+1}, \dots, x_{\bullet,i-1}, \Psi)$ can be computed with two passes through Felsenstein's algorithm, one for the numerator and one for the denominator of Equation 5, each with a different initialization strategy.

When Markov dependence between sites is combined with context-dependent substitution models, it is particularly useful to distinguish between sites of different types. Various *functional categories* of sites might be considered, for example, corresponding to first, second, and third codon positions, and noncoding regions. Alternatively, categories might correspond to different evolutionary rates, or to combinations of functional role and evolutionary rate (Siepel and Haussler 2004). In general, k categories are

described by k phylogenetic models, $\Psi = (\Psi_1, \dots, \Psi_k)$, reflecting the different rates and/or patterns of substitution observed in the different categories. If each site can be assigned a category a priori, with $c(i)$ being the category of the i th site ($1 \leq i \leq L$; $1 \leq c(i) \leq k$), then in the 1st order case ($N = 1$) we have $P(x_\bullet | \Psi) = \prod_{i=1}^L LP(x_{\bullet,i} | \Psi_{c(i)})$. In the general N th order case, different context effects for different categories, as well as different rates and patterns of substitution, can be accommodated similarly, by altering equation 4 so that the probability of each $x_{\bullet,i}$ is conditioned on $\Psi_{c(i)}$. For example, with $N = 2$ and k functional categories, equation 4 would be replaced by:

$$P(x_\bullet | \Psi_1, \dots, \Psi_k) = P(x_{\bullet,1} | \Psi_{c(1)}) P(x_{\bullet,2} | x_{\bullet,1}, \Psi_{c(2)}) \times \prod_{i=3}^L P(x_{\bullet,i} | x_{\bullet,i-2}, x_{\bullet,i-1}, \Psi_{c(i)}). \quad (6)$$

When the categories are not known a priori, it is possible to consider all possible assignments of sites to categories by treating the model as a hidden Markov model, with categories corresponding to states, and emissions corresponding to alignment columns. Such an approach also allows the category of each site to be predicted (Felsenstein and Churchill 1996; Yang 1995; Goldman, Thorne, and Jones 1996; Pedersen and Hein 2003; Siepel and Haussler 2003).

The Markov model presented here is clearly more limited in several respects than the process-based models of Jensen and Pedersen (2000; Pedersen and Jensen 2001). In particular, context effects cannot "cascade" in both directions along a single branch of the tree because inferences are restricted to N -tuples of bases along each branch. Furthermore, because the interdependencies are ignored between the ancestral states associated with overlapping column tuples, there is no globally consistent probabilistic treatment of the latent variables and their interdependencies. Nevertheless, the conditional distributions defined for this $(N - 1)$ st order Markov process are valid probability distributions, and they legitimately combine to produce a valid likelihood: the sum of the probabilities of all alignments is 1 for a given L and n . Most importantly, the model allows efficient computation of likelihoods, and (as will be seen below) fits the data well.

Instead of estimating parameters separately for each functional category under the assumption of independent tuples, parameters can be estimated directly with respect to the entire Markov model (equation 4), using a quasi-Newton algorithm (the EM algorithm described above can no longer be used). It is reasonable to expect, however, that parameter estimates will be similar to those obtained under an assumption of independent tuples. This conjecture is examined experimentally later in this article and will be found to hold reasonably well.

Data

Our noncoding data set was drawn from a 1.8-megabase region of human chromosome 7 containing the CFTR gene and homologous sequences from eight other eutherian mammals (Thomas et al. 2003) (see species in

fig. 4). The sequences were aligned with a new program called TBA (for “Threaded Blockset Aligner”; Blanchette et al. 2004), which was designed specifically for alignment of megabase-sized regions of multiple mammalian genomes. TBA builds a multiple alignment from local pairwise alignments, using a progressive alignment strategy and a predefined phylogenetic tree. It can “thread” a set of locally aligned blocks with respect to one sequence, such that every base of that sequence is represented exactly once, and all bases appear in order. In this case, the tree topology of Figure 4 was used, and the alignment was threaded with respect to the human sequence.

Sites were selected from this alignment corresponding to *ancestral repeats* (ARs)—transposons that appear to have been dispersed, and then become inactive, prior to the divergence of the species in question, and that are believed to have been evolving more or less neutrally since that time. Ancestral repeats were identified in the human sequence, using the program RepeatMasker (<http://ftp.genome.washington.edu/cgi-bin/RepeatMasker/>), and were mapped to the coordinate system of the multiple alignment. The corresponding columns were then extracted and combined into a subalignment, referred to below as the “AR alignment.” Essentially the same set of repeat subfamilies was considered as in recent papers by the Mouse Genome Sequencing Consortium (2002) and Hardison et al. (2003).

The AR alignment was post-processed in various ways to eliminate artifacts from both the alignment procedure and the extraction of columns. To diminish the effects of alignment errors adjacent to indels (due to the placement of gaps to optimize an alignment score), a heuristic was used by which the 3 bases adjacent to each indel were discarded, as were maximal ungapped subsequences shorter than 15 bases in length (discarded bases were replaced by missing data characters—‘N’s). Subsequently, columns having actual bases (non-gaps and non-‘N’s) present in fewer than four species were also removed, so that sites at which two-thirds or more of the species either had gaps or bases bordering gaps were excluded from our analysis. Whenever one or more columns were removed, they were replaced with $N - 1$ columns of missing data (where N is the order of the substitution model to be applied), so that no artificial context was introduced (this “padding” procedure was also used when the AR sites were originally extracted from the complete alignment). The final, cleaned alignment consisted of 162,743 sites, with an average of 5.7 species represented at each site, and a set of padding columns about once every 36 sites. In the end, our results appeared not to be very sensitive to the alignment and post-processing methods (see *Discussion*).

For the coding data set, we used mRNA sequences from various mammalian species that match known genes in the human genome. (It is not yet possible to assemble a data set of coding sites from mammalian genomic DNA with comparable size and species diversity to the non-coding data set described above; the CFTR data set, for example, contains only about 20,000 sites in coding regions.) Using the alignments of nonhuman mRNAs to

the human genome that are available from the UCSC Human Genome Browser (Kent et al. 2002), we selected those genes from a nonredundant set of RefSeq genes (Pruitt and Maglott 2001) having aligned mRNAs from at least two of eight (nonhuman) mammalian species. The same species were used as in the CFTR data set, and only autosomal genes were considered. The best matching mRNA from each species was selected for each gene, with a requirement that this mRNA matched the human genome significantly in no other place. (This rule acted approximately like a reciprocal-best criterion for orthology, but it was somewhat more conservative; it was a more natural choice here because of the asymmetry between the human genome and the nonhuman mRNAs, and because of the redundant nature of the mRNA data.) The coding exons of each RefSeq gene were extracted from the human genome, and re-aligned to the putatively orthologous mRNAs using the T-Coffee program (Notredame, Higgins, and Heringa 2000). The resulting multiple alignments were then cleaned by removing all sites upstream of the start codon and downstream of (and including) the stop codon in the human sequence, all sites with gaps, and (maximal) gapless segments of the alignment shorter than 30 sites in length. In addition, alignments were discarded in which frame-shift indels or premature stop codons occurred in any species (if the anomaly occurred in the last 20% of sites in the gene, only the downstream portion was discarded; the criterion for frame-shift indels was based on the total number of gap characters in each sequence between retained gapless blocks of the alignment, relative to the number in the human sequence). After the cleaning procedure, 2,441 alignments (genes) remained, with an average of 3.4 species per alignment (one of which was always human). Aside from the human genes themselves, this data set is dominated by rodent sequences (2,396 alignments include mouse or rat), and in particular by three-way human/mouse/rat alignments (1,375 alignments); however, cow is fairly well represented (695 alignments), as is pig (433 alignments). Chimp and baboon have the poorest representation, each being present in fewer than 30 alignments. For the purposes of fitting phylogenetic models, all 2,441 alignments were concatenated into a single large alignment consisting of 3,081,993 sites; the species that were absent for each gene were represented in the corresponding columns of this alignment by missing data characters (‘N’s). This alignment has no gaps or stop codons and maintains reading frame completely, with every consecutive triplet of columns representing a column of codons. It is referred to below as the “mRNA alignment.”

Results

This section begins with results for the noncoding data set (the AR alignment), where all sites are treated equally, and then turns to results for the coding data set (the mRNA alignment), where it is important to account for the codon position of each site. In both cases, the likelihoods of various models will first be compared, then the estimates obtained for model parameters will be discussed.

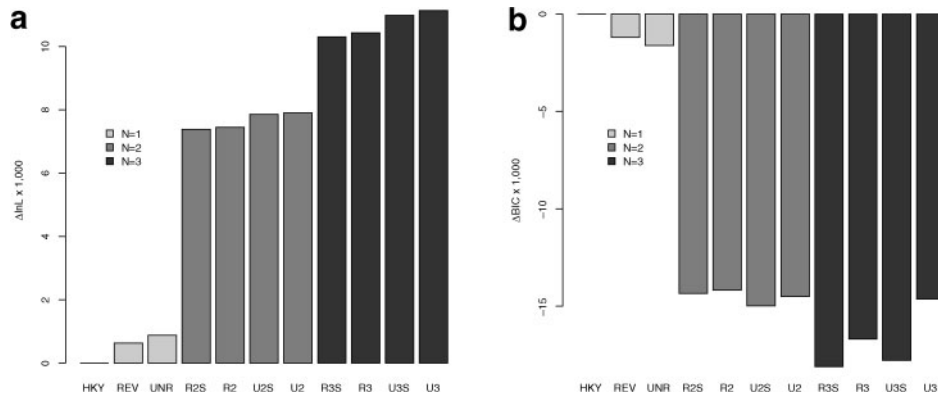


FIG. 1.—*a*. Log likelihoods of various models with respect to the AR alignment (noncoding data), relative to the log likelihood of the HKY model, without rate variation (see table 1). Models with $N = 2$ and $N = 3$ are context dependent. Independent N -tuples of sites were assumed for parameter estimation and likelihood evaluation. *b*. Similar plot showing the Bayesian information criterion (BIC) in place of the log likelihood (BIC = $-2 \cdot \log \text{lik} + m \log L$, where m is the total number of parameters and L is the number of sites in the alignment; here, $L = 162,743$).

Results for Noncoding Data

Various models, with orders from $N = 1$ to $N = 3$, were fitted to the AR alignment, both assuming constant rates across sites and allowing for rate variation (discrete gamma method, $k = 4$ rate categories). Parameters were estimated with the EM algorithm described above, assuming independence of column tuples (with context-dependent models, the sites of the alignment were arbitrarily partitioned into adjacent N -tuples; see below). In all cases, the tree was assumed to have the topology shown later in figure 4, which has been confirmed to be correct by the presence and absence of particular instances of interspersed repeats in the sequences of the CFTR data set (Thomas et al. 2003). Alignment gaps were treated as missing data. Recall that the alignment consisted of $L = 162,743$ sites, and that alignment gaps and ‘N’s accounted for roughly one-third of the characters in each alignment column (on average).

Model Likelihoods

The log likelihoods of all models, under the assumption of constant rates, are shown in Figure 1*a*. Values are plotted relative to the simplest model, HKY. A striking improvement can be seen as the order of the models is increased from $N = 1$ to $N = 2$, and again from $N = 2$ to $N = 3$, indicating strong context effects in the substitution process. Indeed, these improvements far exceed what is achieved by increasing parameter complexity in models of the same order. Nevertheless, most of the improvements among models of the same order do appear to be significant. In several cases, such models can be compared using the standard likelihood ratio test (Huelsenbeck and Rannala 1997), which is applicable between “nested” models, $M_1 \subseteq M_2$; here, $\text{HKY} \subseteq \text{REV} \subseteq \text{UNR}$, and $\text{RNS} \subseteq \text{RN} \subseteq \text{UN} \supseteq \text{UNS}$, for $N = 2, 3$. With all comparable pairs of models, the improvement of the more complex one over the less complex one is statistically significant by this test ($P < 0.0001$), except in the case of U3 versus U3S ($P = 0.2$).

The likelihood ratio test is not applicable for models of different order, and it can be misleading with large data

sets (given enough data, it is possible to obtain small P values for complex models that fit the data only slightly better than simpler counterparts). The Bayesian information criterion (BIC) (Schwartz 1979) is an alternative way of evaluating improvements in likelihood vis-à-vis increases in model complexity, which can be used for non-nested models, and which considers the size of the data set. The BIC strongly supports the use of second-order models over single-nucleotide models, and it strongly supports third-order models over second-order models (fig. 1*b*). Among second-order models, it finds strand symmetry (compare R2S and R2, U2S, and U2) to be a preferable constraint on the rate matrix to reversibility (compare R2S and U2S, R2, and U2), with U2S achieving the best (lowest) score overall. This relationship does not hold for the third-order models, where the cost of the additional parameters of unrestricted models outweighs the resulting improvements in likelihood, and the assumption of reversibility is supported. We note, however, that the BIC is known to tend to overpenalize complex models (Hastie, Tibshirani, and Friedman 2001).

Figure 2 shows, for selected models, the improvements in likelihood obtained by allowing rate variation and overlapping column tuples. Allowing rate variation with this data set makes only a small difference in likelihood (for most models, allowing rate variation is supported by the likelihood ratio test but not by the BIC), and leaves the relative scores of the models essentially unchanged (fig. 2*a*). Apparently, the rate of substitution is not highly variable in these AR sites, consistent with the assumption of neutral evolution. In contrast, a very large increase in likelihood is observed when Markov dependence between columns is introduced—indeed, the improvement over the single-nucleotide models is nearly doubled for dinucleotide models and increases by about half for nucleotide-triplet models, despite no change in the number of parameters of the model (fig. 2*b*). Notice that, if tuples are assumed independent, then context is ignored at every N th boundary between adjacent sites; if Markov dependence is subsequently introduced, then N boundaries are considered in place of each $N - 1$, for a relative gain of $\frac{N}{N-1}$ (2 for dinucleotides and 1.5 for nucleotide triplets).

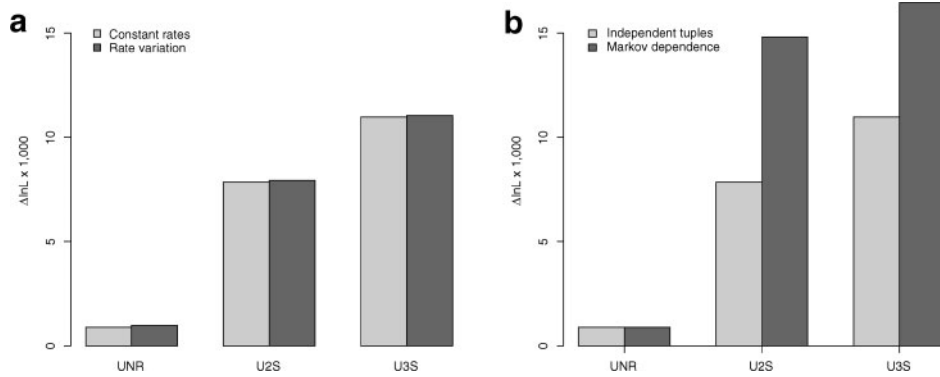


FIG. 2.—The effect on log likelihoods of allowing rate variation (*a*) and Markov dependence between columns (*b*), for selected models (AR alignment). In both plots, the light bars correspond to bars of figure 1*a*, and the dark bars indicate the improvements obtained. All values are relative to the log likelihood of the HKY model without rate variation.

This effect appears in large part to explain the observed improvements, and also explains why the gap between the U2S and U3S models closes substantially when Markov dependence is introduced (fewer boundaries are ignored for U3S when independent tuples are assumed). Note that the parameters of all models were estimated under an assumption of independent tuples, so that the likelihoods in figure 2*b* represent a lower bound on what can be obtained with the Markov-dependent models; direct optimization of the parameters of these models could improve likelihoods further (see below).

Two additional experiments were performed to test whether the increased likelihoods of context-dependent models, with and without Markov dependence of sites, might still somehow be an artifact of the construction of the models. The first experiment was a twofold cross-validation test based on the same data set (the AR alignment). The alignment was divided in half, selected models were trained on approximately the first $L/2$ sites, and likelihoods were evaluated on the second $L/2$ sites ($L =$

162,743); then the procedure was reversed, with training performed on the second half and likelihoods evaluated on the first. The second experiment involved randomizing the columns of an alignment, so that context was effectively erased, then fitting various models to the randomized data, and comparing likelihoods with those obtained from the original alignment. This experiment was performed with the first half of the AR data. The results of these two experiments, shown in figure 3, confirm that the improved likelihoods of the context-dependent models are due to their ability to capture real properties of the biological data, and not to statistical artifacts.

Finally, we tested the conjecture that parameter estimates can reasonably be obtained under an assumption of independent tuples, and then applied in a model that assumes Markov dependence of sites. At the same time, we examined whether it made any difference how a data set was partitioned into independent N -tuples, and whether those tuples were nonoverlapping. We fitted the U2S model to the AR alignment in four different ways: by

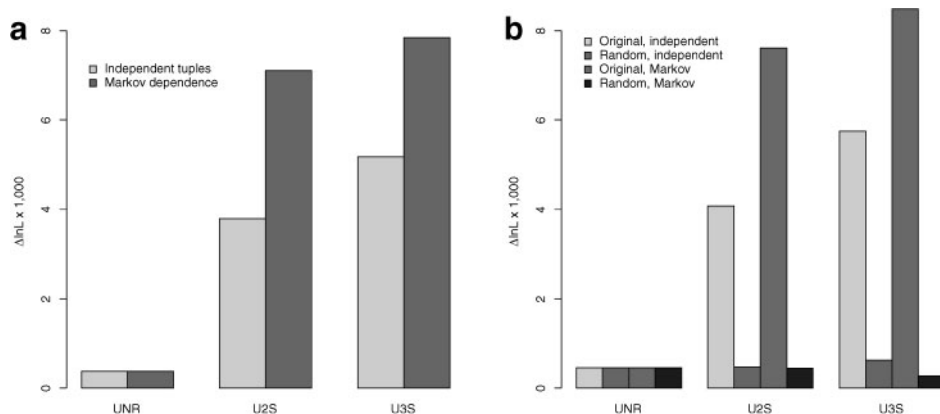


FIG. 3.—Results of the cross-validation (*a*) and column-randomization (*b*) experiments. *a*. The log likelihoods of selected models are shown with respect to a test data set (the second half of the AR alignment), after model parameters had been estimated with respect to a separate training data set (the first half). The relative improvements shown in the previous figures remain essentially unchanged. The other part of the two-fold cross-validation test produced similar results (not shown). *b*. Training log likelihoods are plotted for the first half of the AR alignment (“Original”) and for a version of the same data set in which the columns had been randomly permuted (“Random”). The advantage of the context-dependent models is eliminated with the randomized data. Note the slight likelihood increase for the randomized data under the assumption of independence, and the decrease under the assumption of Markov dependence, which reflect parameter overfitting (models were trained under the independence assumption). This effect should be diminished with larger training sets. All values in both panels are plotted relative to the HKY model.

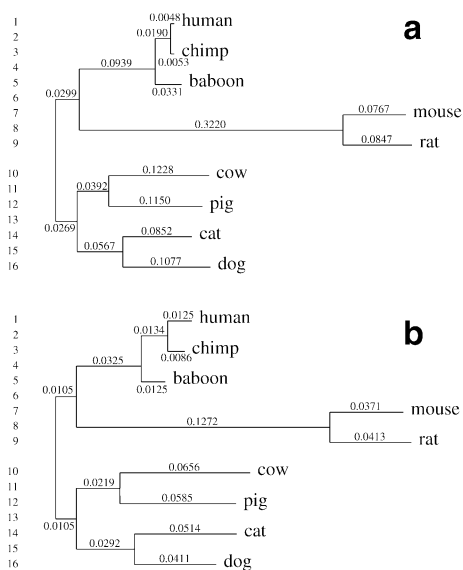


FIG. 4.—The assumed tree topology with branch lengths in substitutions/site as estimated for the AR alignment (*a*) and the mRNA alignment (*b*), under the U3S and R3 models, respectively (with rate variation). *b*. The root of the tree has been arbitrarily placed at the midpoint of the branch between the primates/rodents and the other species (the estimated tree was unrooted). Branch lengths are drawn to scale, separately for each tree. Labels for branches appear to the left of each tree and are used in table 2.

assuming independent tuples (pairs) and training on (1) the “odd” pairs (all adjacent pairs $(x_{\bullet,i}, x_{\bullet,i+1})$ such that i is odd) (2) the “even” pairs, (3) both the odd and even pairs (so that each site was actually considered twice); and (4) by optimizing the full Markov model directly (computationally expensive but possible with $N=2$). The likelihood of each model was then evaluated on the same data under the opposite assumption—that is, Markov dependence in the first three cases and independent tuples in the fourth—and all likelihoods (training and testing) were compared under each assumption. The likelihoods were very similar in all cases (within about 75 units of log likelihood), as were the actual parameter estimates (results not shown).

Parameter Estimates

All parameter estimates discussed in this section were obtained with the EM algorithm, under the assumption of independent tuples. The estimates of branch-length parameters were very similar under all models (those for U3S with rate variation are shown in fig. 4*a*). Indeed, estimates for 2nd and 3rd order models differed from the values shown in figure 4*a* by an average of only 1% to 2%, and estimates for the HKY model without rate variation (the simplest model considered) differed from them by an average of only 5.2%. In general, branch-length estimates increased slightly with the number of parameters in the substitution model, and in particular with the introduction of rate variation, as has been noted in studies of independent-site models (Yang, Goldman, and Friday 1994). Estimates of the shape parameter α of the gamma distribution were very similar for models of

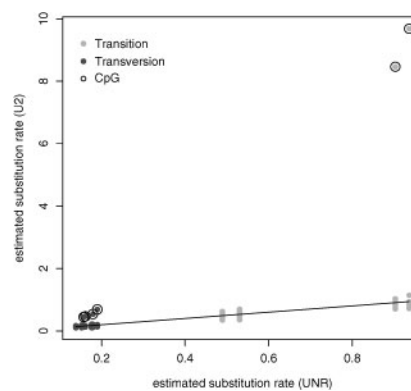


FIG. 5.—Substitution rates for the AR alignment as estimated under the U2 model versus rates estimated under the UNR model (unrestricted with independent sites). Each point represents a non-zero, non-diagonal element in the 16×16 dinucleotide rate matrix in the vertical dimension, and the corresponding element in the 4×4 single-nucleotide rate matrix in the horizontal dimension. The line $y=x$ is shown for reference. Points corresponding to CpG substitutions on either strand are circled (substitutions of the form $CG \rightarrow bG$ or $CG \rightarrow Cd$, where $b, d \in \{A, C, G, T\}$, $b \neq C$, $d \neq G$). Similar plots were obtained for other models, with both $N=2$ and $N=3$. The large differences in estimated transition rates for the UNR model (two rates are ~ 0.5 and two are ~ 0.9) appears to be mostly a consequence of G + C content (about 39% here)—transitions to G and C are estimated to occur at a lower rate than transitions to A and T.

the same order, but they increased slightly with model order (from about 8 for $N=1$ to about 11 for $N=3$). The ratio of the expected transition rate to the expected transversion rate, $\hat{\rho}_{ts}/\hat{\rho}_{tv}$ (Appendix B), was very similar for models of the same order, but it increased slightly with model order (2.03 for $N=1$, 2.12–2.14 for $N=2$, and 2.16–2.17 for $N=3$).

More interesting are the estimates of specific rate-matrix parameters, which provide insight about context-dependent substitution in noncoding DNA. For context-dependent models, the elements of the estimated rate matrix \mathbf{Q} (which, in the case of reversible models, also incorporate the equilibrium frequencies) spanned a wide range of values (e.g., 0.05–10.9 for the U3S model) and tended to cluster into three main groups, corresponding to transversions, transitions, and CpG transitions. CpG transversions ($CG \rightarrow AG$ and $CG \rightarrow GG$, and their reverse complements) occurred at rates comparable to those of non-CpG transitions. Comparisons of context-dependent models with independent-site models (fig. 5) confirmed that the most pronounced context effects corresponded to CpG transitions, although CpG transversions also occurred at a significantly higher rate than expected (note that the mechanism behind CpG transversions is believed to be different from that behind CpG transitions; see, e.g., Blake, Hess, and Nicholson-Tuell 1992). CpG effects also appear to explain, at least in part, the somewhat poor fit of reversible, context-dependent models (as described above). When estimates of substitution rates under the U3 and U3S models are compared (fig. 6*a*), they are seen to be very closely correlated, indicating that the constraint of strand symmetry does not prohibit a near-optimal version of the matrix from being obtained. A similar comparison of the U3 and R3 models, however (fig. 6*b*), shows substantially poorer agreement for certain classes of

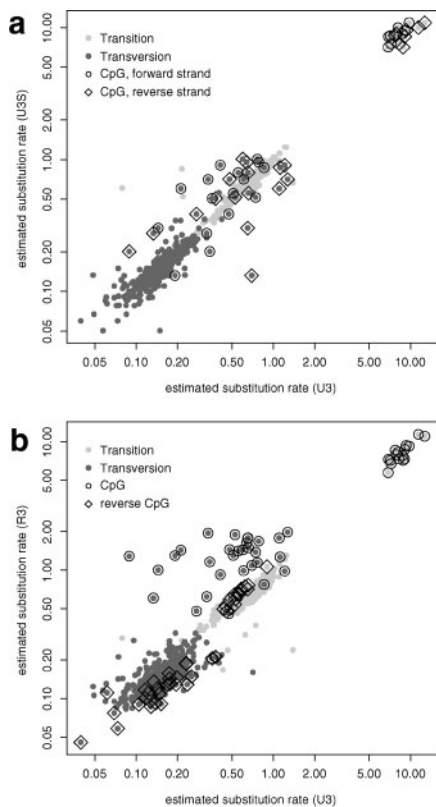


FIG. 6.—Substitution rates estimated for the AR alignment under the U3 and U3S models (*a*), and under the U3 and R3 models (*b*). *a*. CpG substitutions ($aCG \rightarrow abG$ or $CGa \rightarrow bGa$, where $a, b \in \{A, C, G, T\}$ and $b \neq C$) are indicated by circled points and their strand-symmetric counterparts by points enclosed in diamonds. *b*. CpG substitutions on either strand ($aCG \rightarrow abG$, $CGa \rightarrow bGa$, $aCG \rightarrow aCd$, $CGa \rightarrow Cda$, where $b \neq C$, $d \neq G$) are indicated by circles, and the reverse substitutions are indicated by diamonds. In both *a* and *b* the tendency of the rates to group into (non-CpG) transitions, (non-CpG) transversions, and CpG transitions is evident, with CpG transversions clustering with non-CpG transitions.

substitutions, particularly CpG transversions ($r = 0.55$ for CpG transversions under U3 vs. U3S and $r = 0.36$ for CpG transversions under U3 vs. R3). The rates of these transversions appear to be systematically overestimated under R3, and the rates in the reverse direction appear to be underestimated.

It is worth noting that our results do not reflect the transcription-associated mutational asymmetry recently reported by Green et al. (2003), based on an analysis of the same sequence data, although a substantial portion of the sites we have considered are believed to be transcribed. Three of the nine genes in the region in question, however, are located on the opposite strand from the other six, and so our data set contains a mixture of bases not only that are and are not transcribed, but that correspond to the transcribed and nontranscribed strands. This is presumably why strand asymmetries are not apparent from a comparison of the UN and UNS models.

The estimated rates for the U3S model were compared to rates of context-dependent substitution in human pseudogenes estimated by Hess, Blake, and Blake (1994), in an expansion of the study by Blake, Hess, and Nicholson-Tuell (1992). The reported “corrected relative

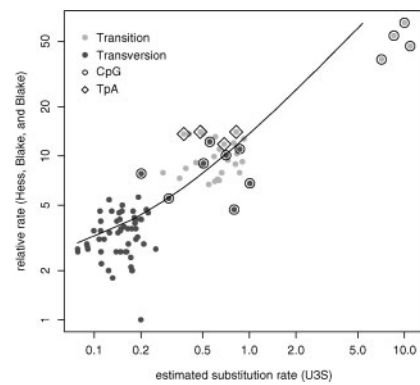


FIG. 7.—Substitution rates for the AR alignment as estimated under the U3S model versus relative rates in human pseudogenes reported by Hess, Blake, and Blake (1994). Points corresponding to CpG substitutions are circled, and points corresponding to TpA substitutions are enclosed in diamonds. Strand symmetry has been assumed for both sets of estimates, so only one point appears for each strand-symmetric pair of substitutions.

rates” were used, which are ratios with respect to the slowest estimated rate, with an upward correction applied to CpG substitutions. The results of the comparison are shown in figure 7. The two sets of estimates are remarkably consistent overall, considering the difference in methods and our use of ARs from a single, local region of the genome instead of pseudogenes from various chromosomes. The largest disagreements correspond to CpG transitions, for which we estimate considerably higher rates (1.8–2.8 times higher than would be predicted by the regression line of figure 7). These disagreements may reflect real differences in the two data sets—e.g., due to differences in the methylation patterns in AR sites versus pseudogenes—but it is likely that they are at least partially a consequence of a non-negligible rate of multiple substitutions per site, or of ascertainment bias in Hess, Blake, and Blake’s (1994) methods (their selection procedure, which was designed to eliminate cases of multiple substitutions per site, may have been biased against sites with CpG transitions).

No simple explanations emerged for the variation in the estimates of non-CpG rates, which was considerable, both among transitions and among transversions (non-CpG transition rates for U3S had mean 0.653 and s.d. 0.203 and non-CpG transversion rates had mean 0.161 and s.d. 0.057). Indeed, several reported trends were not supported by our parameter estimates. For example, Hess, Blake, and Blake (1994) observed high rates of $TA \rightarrow CA$ transitions, but we did not find these rates to be unusual (fig. 7). Our estimates also did not show a significant dependency of the transition/transversion ratio on the A + T content of neighboring bases: the expected proportion of substitutions that are transversions, based on our estimated rate matrix, increased only slightly with the number of flanking A + T bases (under the U3 model, from 0.275 with 0 flanking A + T bases, to 0.290 with 1, and to 0.349 with 2), and this increase seemed to be largely a consequence of the CpG effect (when CpG substitutions were discarded the expected numbers were

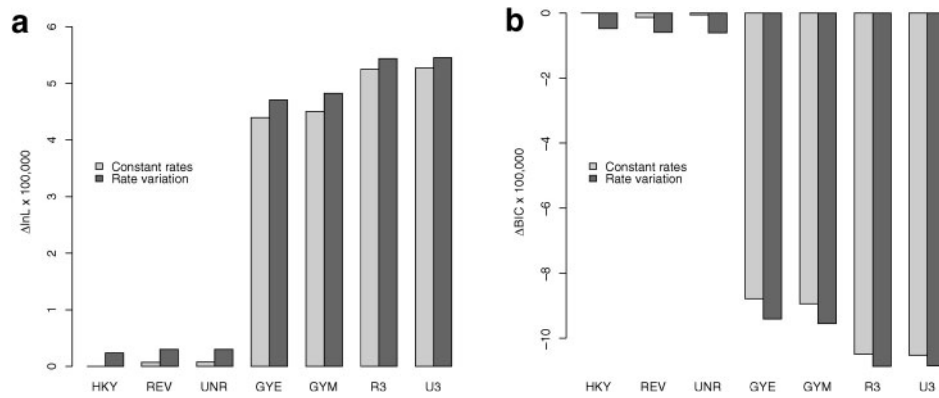


FIG. 8.—*a*. Log likelihoods of various models with respect to the mRNA alignment (coding data), with and without rate variation. N -tuples of sites were assumed independent in parameter estimation and likelihood evaluation. *b*. Similar plot showing the BIC in place of the log likelihood (see legend of figure 1). The number of sites here is $L = 3,081,993$.

0.334 for 0 A + T bases, 0.320 for 1, and 0.349 for 2). (We note that the proportions we estimated *including* CpG substitutions are comparable to those reported for primates by Zhao and Boerwinkle (2002) and Yang, Chen, and Li (2002), who appear not to have corrected for the CpG effect.) Our estimated rates also did not reflect the tendency observed in chloroplast genomes for a higher proportion of transversions when a 5' pyrimidine is present (Morton, Oberholzer and Clegg 1997). As has been noted for chloroplast genomes (Morton, Oberholzer, and Clegg 1997), the pattern of context-dependent substitution in the noncoding DNA of mammals appears to be complex, and is not easily explained in terms of a small number of well-defined phenomena (such as the CpG effect and the transition/transversion bias).

Results for Coding Data

A different set of models was fitted to the mRNA alignment, this time including the GYE and GYM versions of Goldman and Yang's (1994) codon model (Appendix B), as well as the HKY, REV, UNR, R3, and U3 models. Second-order models and strand-symmetric models were not considered. All models were fitted with and without rate variation. The EM algorithm was used, assuming independent column tuples (here corresponding to codons, for the third-order models), except for Goldman and Yang's models, which were fitted using the codeml program in the PAML package (Yang 1997). The same tree topology was assumed as in the previous section. Recall that the mRNA alignment consisted of 993 sites, $L = 3,081$, with 3.4 species represented per site; sites with alignment gaps had been removed.

For the R3 and U3 models, the effect of assuming Markov dependence between columns was also examined. In this case (unlike in the previous section), three separate functional categories were considered, corresponding to the 1st, 2nd, and 3rd codon positions, and a separate third-order model was fitted to the sites of each category, using the EM algorithm. Thus, the third codon position model was treated like an ordinary codon model, but the first and second position models were trained on triples that straddled adjacent codons, so that context effects across

codon boundaries were considered. All three models were incorporated into the likelihood calculation, as shown in equation 6.

Model Likelihoods

Figure 8 shows the log likelihoods and BIC scores of all models, with and without rate variation. The codon models are seen to fit the data overwhelmingly better than the single-nucleotide models, indicating pronounced context-dependence among sites of the same codon. The R3 and U3 models, however, performed substantially better than Goldman and Yang's model; apparently, the pattern of codon substitution is complex, and is characterized only approximately by these simply parameterized models. The use of a physicochemical distance matrix (GYM) produced a relatively small improvement over the naive assumption of equal distances between amino acids (GYE), supporting Yang, Nielsen, and Hasegawa's (1998) conclusion that physicochemical distances and substitution rates are not strongly correlated (see below). We expect that the R3 and U3 models would improve similarly on other existing codon models, which have been reported to perform roughly as well as Goldman and Yang's model (Schadt and Lange 2002).

Allowing for rate variation made a large difference for all models, but a larger difference for single-nucleotide models and Goldman and Yang's codon model than for R3 and U3; the reason is probably that rate variation is being used to compensate for an oversimplified representation of substitution patterns (estimates of the shape parameter α were about 0.40 under the single-nucleotide models, 0.95 under GYE and GYM, and 1.4 under R3 and U3).

The U3 model outscored the R3 model by about 2,000 units of log likelihood (with and without rate variation), a sufficient improvement to reject the hypothesis of reversibility under the likelihood ratio test. The R3 and U3 models are almost equal according to the BIC, however, and indeed, R3 is slightly preferable in the case with rate variation. Despite some evidence against reversibility of codon substitution, it appears to be a rea-

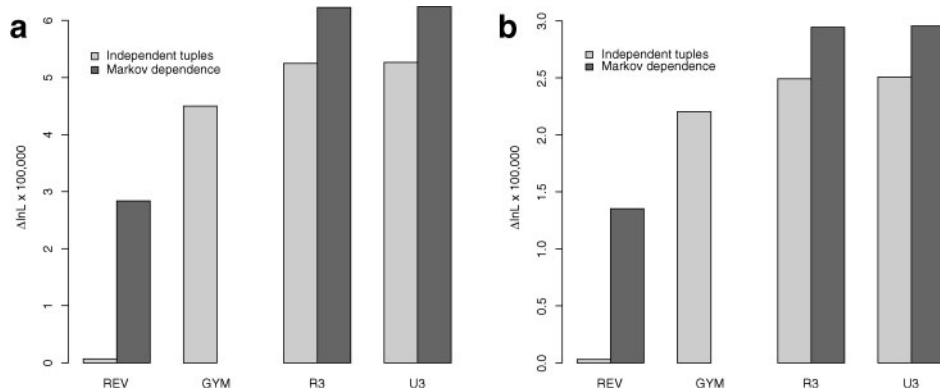


FIG. 9.—*a*. Log likelihoods of various models with respect to the mRNA alignment, with and without assuming Markov dependence of sites. All values are relative to the log likelihood of the HKY model without rate variation. The Markov model is not compatible with Goldman and Yang's model, which can only be applied to in-frame codons. For the Markov-dependent REV model, separate models were fitted to sites of each codon position, and their log likelihoods were combined. *b*. Similar plot showing cross-validation results, with models trained on the first half of the mRNA alignment, and likelihoods evaluated on the second half. (The training likelihood for the second half of the alignment is shown for the GYM model, because software was not available to train and test on different data sets; this value represents an upper bound on what would be obtained under cross-validation.) The relative scores of the models remain essentially unchanged.

sonable modeling assumption, especially when its mathematical convenience is taken into consideration.

When Markov dependence between sites was introduced with the U3 and R3 models (fig. 9*a*), another large improvement in likelihoods occurred, indicating that context effects that cross codon boundaries are important (see also Pedersen, Wiuf, and Christiansen 1998). This improvement holds up well under cross-validation (fig. 9*b*). It remains to be determined to what extent the CpG effect is responsible for the observed improvements in likelihood (e.g., because of synonymous transitions at CpG), and in general, how similar the context effects in coding DNA (particularly at unconstrained sites) are to the ones observed in putatively neutral DNA.

Parameter Estimates

The estimated branch lengths based on the mRNA alignment were similar for all models, but they were less consistent than the estimates for the noncoding data set. Results for selected models are shown in Table 2, with branch-length labels defined in figure 4*b*. (Throughout this section, estimates are presented for the versions of R3 and U3 that were trained like ordinary codon models.) The estimates for single-nucleotide models differed considerably at several branches from those for the codon models, and Goldman and Yang's model produced estimates at a few branches that differed significantly from those of R3 and U3. Most of the major differences, however, appeared in the subtree of the primates, and they should be interpreted with caution, because of sparse data for chimp and baboon, and because noise in the mRNA sequences will have a disproportionate effect on short branches. In general, these results suggest that simply parameterized codon models, and even independent-site models with rate variation, may be adequate for branch length estimation, despite the improved fit of richer models.

The relative proportions of branch lengths estimated for the noncoding and coding data sets are fairly similar overall (fig. 4), given the very different evolutionary pressures in coding and noncoding DNA, and the fact that

the mRNA data are drawn from around the genome whereas the AR data come from a single local region. We interpret this general congruence as an indication that we have obtained a reasonable approximate characterization of the relative average rates of substitution on the branches of the tree, although the absolute values of the estimates are subject to many possible biases (for example, related to alignment). Interestingly, the ratio of the distances (in substitutions/site) to mouse and human from their common ancestor is estimated to be between about 2.8:1 (mRNA data, R3 with rate variation) and 3.4:1 (AR data, U3S with rate variation), which is considerably higher than the estimate of 2:1 given in the recent analysis of the mouse genome (Mouse Genome Sequencing Consortium 2002). Ratios similar to ours were estimated by Cooper et al. (2003).

Even more than with the noncoding data, a wide range was observed in the estimated substitution rates (from 0.001 to 4.63 for different codon substitutions under the U3 model). The estimated matrices for U3 and R3 were comparable. The estimated transition/transversion ratio, $\hat{\rho}_s/\hat{\rho}_{iv}$, was similar for all codon models (e.g., estimates without rate variation were 2.17 for GYE and GYM and 2.15–2.16 for R3 and U3), and it was somewhat lower under the independent-site models (1.97 for HKY, REV, and UNR). Estimates of the ratio of the rates of synonymous and nonsynonymous substitution, $\hat{\rho}_s/\hat{\rho}_a$ (Appendix B), were also similar under all codon models (2.80–2.82 for GYE and GYM, 2.80 for R3, and 2.72 for U3, without rate variation).

The amino acid substitution rates induced by the estimates for the R3 and U3 models (see Appendix B) showed good agreement with empirical substitution matrices, such as the one of Jones, Taylor and Thornton (1992; fig. 10). It appears that amino acid substitution rates can be estimated implicitly from nucleotide sequences, given enough data and a sufficiently rich parameterization of the substitution model. In contrast, the rates used by the GYM model, which are based on the physicochemical distances of Miyata, Miyazawa, and Yasunaga (1979;

Table 2
Branch Lengths Estimated for the mRNA Alignment Under Selected Models (substitutions/site)

Branch	REV	GYM	R3	REV + Γ	GYM + Γ	R3 + Γ
1	0.0078	0.0103	0.0120	0.0074	0.0116	0.0125
2	0.0073	0.0043	0.0130	0.0058	0.0044	0.0134
3	0.0327	0.0091	0.0088	0.0257	0.0077	0.0086
4	0.0324	0.0372	0.0262	0.0396	0.0467	0.0325
5	0.0405	0.0200	0.0118	0.0398	0.0199	0.0125
6	0.0079	0.0088	0.0088	0.0086	0.0114	0.0105
7	0.0323	0.0341	0.0340	0.0346	0.0387	0.0371
8	0.0811	0.1004	0.1004	0.1078	0.1398	0.1272
9	0.0350	0.0375	0.0375	0.0382	0.0431	0.0413
10	0.0489	0.0564	0.0563	0.0577	0.0699	0.0656
11	0.0151	0.0180	0.0185	0.0189	0.0241	0.0219
12	0.0468	0.0526	0.0526	0.0521	0.0621	0.0585
13	0.0079	0.0088	0.0088	0.0086	0.0114	0.0105
14	0.0512	0.0577	0.0462	0.0552	0.0604	0.0514
15	0.0208	0.0233	0.0281	0.0225	0.0284	0.0292
16	0.0366	0.0405	0.0366	0.0410	0.0477	0.0411

NOTE.—Models are discussed in table 1. Branch numbers are defined in figure 4*b*. Γ indicates discrete gamma model for rate variation ($k = 4$ categories).

Appendix B), showed much weaker agreement with the induced rates of R3 and U3. These results support the conclusion of Yang, Nielsen, and Hasegawa (1998) that substitution rates and physicochemical distances are not related by a simple mathematical function, and they suggest that the use of physicochemical distances is a weakness of current codon models.

Discussion

Allowing for context-dependent substitution makes phylogenetic inference considerably harder: when the assumption of site-independence is relaxed, a Markov random field arises, and standard methods can no longer be used for likelihood computation and parameter estimation (Jensen and Pedersen 2000). We have shown, however, that context-dependent substitution can be handled in an approximate way with some simple extensions of codon models, which themselves are direct extensions of Felsenstein's original framework. Our models (like codon models) require the assumption of certain limitations on the interdependence between sites (see below), but in return, they allow for exact inference without too much additional cost in computation. (It should be noted that our models span a wide range of computational demand, from the more simply parameterized second-order models, which are practical for use in ordinary phylogenetic analysis, to richly parameterized third-order models, which may be too costly for some purposes—see Appendix C.) The new models appear to fit real biological data substantially better than independent-site and codon models, in both coding and noncoding regions.

The improvements we observe appear to derive from three more or less separate properties of our models: (1) their ability to capture context-effects within independent N -tuples of sites; (2) their ability to allow these N -tuples to overlap; and (3) their rich parameterization of the substitution process. Judging by cases in which some apply but others do not, these three properties have roughly equal importance, and their effects are approximately additive.

When N -tuples are assumed independent (property 1), as in codon and RNA models, standard methods can be used for likelihood computation. In addition, parameter estimation can be accomplished efficiently with an EM algorithm, which performs much better than general-purpose optimization algorithms on parameter-rich models and large data sets. Our models allow N -tuples to overlap (property 2) by assuming Markov dependence between sites. Likelihood computations are performed simply and efficiently, using a two-pass extension of Felsenstein's algorithm. Expectation maximization cannot be used to fit the Markov-dependent models directly, but parameter estimates based on an assumption of independent sites appear to be close to optimal.

The importance of a rich parameterization of the substitution process (property 3) indicates that the patterns of context-dependent substitution are complex in both coding and noncoding regions. In noncoding regions, the CpG effect is the single, strongest, clearly identifiable example of context-dependence, but it does not appear to be sufficient to explain the advantage of context-dependent models; judging by the wide variation in estimates of substitution rates, many more subtle effects also occur, and these may be significant in combination. The improvements in goodness of fit that occur when second-order models are replaced by third-order models suggest that important context effects are occurring at the level of nucleotide triplets; however, it is also true that the third-order models provide a better approximation of the actual process of context-dependent substitution (see below), so these increases should be interpreted with caution. Additional work will be required to sort out which are the most important context effects, and whether simpler parameterizations of context-dependent rate matrices can be justified.

In coding regions, a richly parameterized substitution model allows for significantly higher likelihoods than those of Goldman and Yang's codon model, and by extension, other codon models. Goldman and Yang's

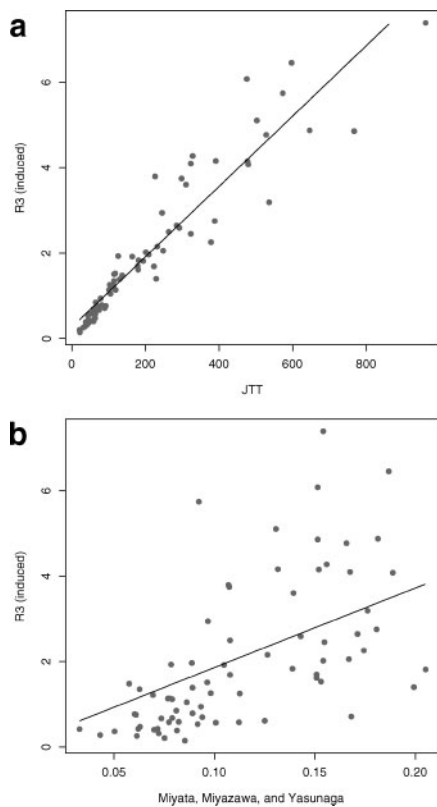


FIG. 10.—*a*. Induced amino acid substitution rates estimated for the mRNA alignment under R3 (without rate variation) versus rates based on the JTT substitution matrix (Jones, Taylor, and Thornton 1992). The version of the JTT matrix that is distributed with the PAML package (Yang 1997, version 3.13) was used, with rates derived from counts using the method of Yang and Kumar (1996). For a proper comparison, each element of the induced matrix has been divided by the equilibrium frequency of the target residue. Points are not shown for induced rates of zero that result from all encoding pairs of codons being different by more than one nucleotide (there are 115 such pairs of amino acids, with a mean JTT score of 26.4). The displayed line is based on a linear regression ($r^2 = 0.88$). Results are similar for the U3 model, and for the Dayhoff (Dayhoff, Schwartz, and Orcutt 1978) and WAG (Whelan and Goldman 2001) substitution matrices (not shown). *b*. Similar plot with respect to the physicochemical distances of Miyata, Miyazawa, and Yasunaga, transformed by the function used in the GYM model ($\gamma_1 \exp[-\gamma_2 d_{c_1, c_2} / d_{\max}]$; see Appendix B), with the values of scaling constants γ_1 and γ_2 that were estimated from the data. The observed correlation is considerably weaker than in panel *a* ($r^2 = 0.73$).

model is appealing in its use of a small number of parameters, each with a clear biological meaning, but it appears not to characterize the pattern of codon substitution as accurately as our more empirical models (it is important to note, however, that its characterization generally seems adequate for accurate estimation of branch lengths and quantities such as ρ_{ts}/ρ_{tv} and ρ_s/ρ_a). The most serious problem seems to be with parameterizing non-synonymous substitution rates in terms of physicochemical distances between amino acids (Yang, Nielsen, and Hasegawa 1998). An alternative would be to use empirically derived amino acid substitution matrices (such as the JTT or Dayhoff matrices), but these matrices are estimated in a way that is not consistent with the assumptions of the continuous-time Markov models used

in phylogenetic models (Goldman and Yang 1994). At the cost of additional parameters, our models allow empirical estimation of amino acid rates within this continuous-time Markov framework, and at the same time capture phenomena at the DNA level, such as the transition/transversion bias and the pattern of synonymous substitution. Thus, to an extent, they combine the benefits of the two competing types of models that Yang, Nielsen, and Hasegawa (1998) called “mechanistic” (codon models) and “empirical” (models applied at the amino-acid level, using empirical substitution matrices). On the other hand, our models rely on a large number of parameters—larger than in any phylogenetic model currently in wide use—which increases the computational burden of parameter estimation and implies that large data sets to obtain accurate estimates. The first of these problems, we argue, can be surmounted with better algorithms and faster hardware, and the second will be partially addressed by the large quantities of data generated by the genome sequencing projects. Nevertheless, models with hundreds of parameters clearly will still be inappropriate for some applications. It may be possible in some cases to use a model with a smaller number of parameters (such as Goldman and Yang’s), in combination with amino-acid substitution rates that have been estimated on a larger data set, using a less constrained model (such as R3).

Although our models improve on the goodness of fit of independent-site and codon models, the question remains open of how they would compare to models that more faithfully describe the true process of context-dependent substitution, such as those of Jensen and Pedersen (2000; Pedersen and Jensen 2001). As discussed above, our models fail to capture the interdependence of the ancestral states associated with overlapping N -tuples, and thus they do not give a consistent probabilistic treatment of the latent variables. This deficiency could be rectified by defining the interactions between all variables, latent and observed, using a general graphical model (Jordan 1999). The result would be a kind of discrete Markov random field with loops of dependency. Although no efficient method exists for exact likelihood calculation from such models, there are efficient approximate algorithms (Murphy, Weiss, and Jordan 1999; Wainwright, Jaakkola, and Willsky 2003; Yedidia, Freeman, and Weiss 2000). Another issue is that our models do not allow context effects to “cascade” outside the boundaries of each N -tuple along an individual branch of the tree, as is allowed in the process-based models of Jensen and Pedersen. The importance of this limitation will decrease as the order of a model increases, but it may be non-negligible when $N = 2$ or $N = 3$; it could be assessed by comparing our models with process-based models using MCMC methods. In the end, exact inference with an approximate model (as described in this article) will have to be weighed against approximate inference with a (more) exact model, considering both the goodness of fit of the models and the computational burden of parameter estimation. It is likely that different approaches will be appropriate for different applications.

The usual caveats about modeling assumptions and alignment accuracy apply to all parameter estimates presented in this article. The assumption that the substitution process is homogeneous and stationary, in particular, undoubtedly does not hold in reality, e.g., due to regional variations in G + C content or differences between lineages in the pattern of substitution (G + C content in our AR data set from the CFTR region is relatively uniform, both across sites and across species, but it varies considerably in our genome-wide mRNA data set). It is not yet clear what effect violations of these assumptions may have. Aspects of this question could be examined by relaxing the assumption of homogeneity in the manner of the model of Yang and Roberts (1995). The coding sequences can be aligned with fairly high confidence, but the alignment of noncoding sequences may introduce certain biases. Experiments with alternative alignments—one for which the post-processing steps were omitted and one produced by the MAVID program (Bray and Pachter 2003)—indicated that the absolute values of branch-length estimates are somewhat sensitive to the alignment and cleaning procedure, but relative branch lengths are fairly robust, as are estimates of context-dependent substitution rates (results not shown). The parameter estimates for the coding data set could conceivably be influenced by erroneous assignments of orthology, or by the use of mRNA (rather than genomic) sequence data.

We have focused on using context-dependent phylogenetic models to estimate the pattern and rates of substitution on the branches of a tree of known topology, but their broader role in phylogenetic modeling is worth considering. In simulation studies concerned with the accuracy of topology reconstruction, maximum-likelihood methods (assuming independent sites) were found to be fairly robust against violations of site independence, but they did have somewhat diminished statistical power in the presence of such violations (Schöniger and von Haeseler 1995). Thus, it is reasonable to expect that context-dependent models will not dramatically improve estimates of topologies, but they might be helpful in certain cases, for example, with hard-to-resolve branchings near the root of the tree. On the other hand, consideration of context effects could lead to significant improvements in cases where phylogenetic models are used to discriminate between different types of sites, such as in gene finding (Pedersen and Hein 2003; Siepel and Haussler 2004) and secondary structure prediction (Goldman, Thorne, and Jones 1996). In addition, context-dependent models may be helpful in understanding the evolution of isochores, in which the CpG effect has been suggested to play an important role (Fryxell and Zuckerkandl 2000).

Supplementary Material

The AR and mRNA alignments and complete descriptions of parameter estimates for all models are available at www.cse.ucsc.edu/~acs/context/. The software used for parameter estimation is available from the authors upon request (acs@soe.ucsc.edu).

Appendix A: Felsenstein's Algorithm

Let u be a node in the tree, and let $x_{u,i}$ denote the observed characters at the leaves beneath u at site i , $x_{u,i} = \{x_{v,i} | v \in \mathcal{L} \text{ and } v \text{ is a descendant of } u\}$. Felsenstein's algorithm computes $P(x_{u,i} | X_{u,i} = a)$ for all nodes u and characters a , using the following recursive rule (here v and w are the children of u):

$$P(x_{u,i} | X_{u,i} = a) = \begin{cases} I(x_{u,i} = a) & \text{if } u \in \mathcal{L} \\ \sum_{b \in \Sigma} P(b | a, \beta_v) P(x_{v,i} | X_{v,i} = b) \\ \quad \times \sum_{c \in \Sigma} P(c | a, \beta_w) P(x_{w,i} | X_{w,i} = c) & \text{if } u \in \mathcal{I}, \end{cases} \quad (\text{A.1})$$

where I is the indicator function. The total probability of the column is obtained as $P(x_{\bullet,i} | \Psi) = \sum_a \pi_a P(x_{\bullet,i} | X_{\bullet,i} = a)$. Felsenstein's algorithm is a special case of what is known in the graphical models literature as the "elimination" algorithm; it is closely related to the "peeling" algorithm from statistical genetics, and the "forward" algorithm for hidden Markov models (Jordan, 2004). The algorithm requires $O(nLd^2)$ time to compute $P(x_{\bullet,i} | \Psi)$, for n taxa, an alignment of length L , and an alphabet of size d .

Felsenstein's algorithm can be adapted easily to accommodate missing data. If a random variable $X_{u,i}$, corresponding to a leaf u , is somehow unobserved (as when alignment character $x_{u,i} = \text{'N'}$), then the algorithm can be made to sum over all possible values of $x_{u,i}$ by changing the base case of the recursion to $P(x_{u,i} | X_{u,i} = a) = 1$ (for all a). The resulting algorithm has the desirable property of returning a probability $P(x_{\bullet,i} | \mathbf{Q}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\beta})$, for a column $x_{\bullet,i}$ with missing data, that is equal to the probability $P(x'_{\bullet,i} | \mathbf{Q}, \boldsymbol{\pi}, \boldsymbol{\tau}', \boldsymbol{\beta}')$, for the induced subtree $(\boldsymbol{\tau}', \boldsymbol{\beta}')$ of $(\boldsymbol{\tau}, \boldsymbol{\beta})$ and a version $x'_{\bullet,i}$ of $x_{\bullet,i}$ that has the missing elements removed. Thus, the probability of a column $x_{\bullet,i}$ with missing data does not depend on the branches leading to the "missing leaves," and when fitting a model, the column $x_{\bullet,i}$ has no bearing on such branches. Alignment gaps are sometimes treated as missing data.

Appendix B: Context-Dependent Models

General Nth-Order Models

Let $\delta(a, b)$ be the number of characters by which tuples $a \in \Sigma^N$ and $b \in \Sigma^N$ differ. We define the general unrestricted substitution model UN, in terms of free parameters of the form $Q_{a,b}$, as:

$$q_{a,b} = \begin{cases} Q_{a,b} & \text{if } \delta(a, b) = 1 \\ 0 & \text{if } \delta(a, b) > 1 \\ -\sum_{b \in \Sigma^N - \{a\}} q_{a,b} & \text{if } \delta(a, b) = 0. \end{cases} \quad (\text{B.1})$$

This model has $N(d-1)d^N$ free parameters, where d is the size of the alphabet in question. The general reversible substitution model RN is the same as UN except for the constraint that $\pi_a q_{a,b} = \pi_b q_{b,a}$, and the strand-symmetric models UNS and RNS are versions of UN and RN, respectively, such that $q_{a,b} = q_{\bar{a}, \bar{b}}$ if \bar{a} and \bar{b} are the

reverse complements of a and b , respectively. Notice that R1 is equivalent to REV, and U1 to UNR.

Once a maximum-likelihood estimate $\hat{\Psi}$ of an N th-order model has been obtained, the expected rates of particular types of substitutions can easily be obtained from the estimated rate matrix, \mathbf{Q} . For example, the transition rate is given by

$$\hat{\rho}_{ts} = \sum_{a,b \in \Sigma^N} \pi_a \hat{q}_{a,b} I(a \leftrightarrow b \text{ is a transition}) \quad (\text{B.2})$$

(Note that equation B.1 ensures that every non-zero element of \mathbf{Q} corresponds to a single-nucleotide substitution.) Because of the scaling constraint on the matrix, the transversion rate is simply $\hat{\rho}_{tv} = N - \hat{\rho}_{ts}$, and the transition/transversion ratio is $\hat{\rho}_{ts}/\hat{\rho}_{tv} = \hat{\rho}_{ts}/(N - \hat{\rho}_{ts})$. The invariance property of maximum likelihood estimators ensures that estimates obtained in this way are maximum likelihood estimates (Goldman and Yang 1994).

Codon Models

A *codon model* is a third order phylogenetic model $\Psi = (\mathbf{Q}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\beta})$, defined with respect to a DNA alphabet, such that the equilibrium frequency of each stop codon is zero. A codon model Ψ has the property of inducing an amino acid model Ψ' in the following way. Let $\Psi' = (\mathbf{Q}', \boldsymbol{\pi}', \boldsymbol{\tau}, \boldsymbol{\beta})$, with \mathbf{Q}' and $\boldsymbol{\pi}'$ defined with respect to an amino acid alphabet, Σ' . Let $\alpha(c)$ either indicate the amino acid a encoded by codon c ($a \in \Sigma', c \in \Sigma^3$) or have a null value if c is a stop codon. Then the elements of $\boldsymbol{\pi}'$ are given by $\pi'_a = \sum_{c: \alpha(c) = a} \pi_c$, and the elements of \mathbf{Q}' by $q'_{a_1, a_2} = \sum_{c_1: \alpha(c_1) = a_1} \sum_{c_2: \alpha(c_2) = a_2} \pi_{c_1} q_{c_1, c_2}$ (Yang, Nielsen, and Hasegawa 1998). If Ψ is reversible then Ψ' will also be reversible.

The codon model of Goldman and Yang (1994), as adapted by Yang, Nielsen and Hasegawa (1998), is defined by the following parameterization of \mathbf{Q} (here denoted $\mathbf{Q} = \{q_{c_1, c_2}\}$):

$$q_{c_1, c_2} = \begin{cases} \pi_{c_2} \cdot \gamma_1 \exp[-\gamma_2 \frac{d_{\alpha(c_1), \alpha(c_2)}}{d_{\max}}] & \text{if } \delta(c_1, c_2) = 1, c_1 \leftrightarrow c_2 \text{ transversion} \\ \kappa \pi_{c_2} \cdot \gamma_1 \exp[-\gamma_2 \frac{d_{\alpha(c_1), \alpha(c_2)}}{d_{\max}}] & \text{if } \delta(c_1, c_2) = 1, c_1 \leftrightarrow c_2 \text{ transition} \\ 0 & \text{if } \delta(c_1, c_2) > 1 \\ -\sum_{c_2 \in \Sigma^N - \{c_1\}} q_{c_1, c_2} & \text{if } \delta(c_1, c_2) = 0, \end{cases} \quad (\text{B.3})$$

where $\mathbf{D} = \{d_{a_1, a_2}\}$ ($a_1, a_2 \in \Sigma'$) is a symmetric amino-acid distance matrix (with $d_{a,a} = 0$ for all $a \in \Sigma'$), d_{\max} is the maximum value in \mathbf{D} , and γ_1, γ_2 are scaling parameters. The model assumes that \mathbf{D} is defined a priori. Goldman and Yang originally used the physicochemical distances of Grantham (1974), but Yang, Nielsen, and Hasegawa (1998) subsequently compared several different physicochemical distance measures and found the one of Miyata, Miyazawa, and Yasunaga (1979) to perform best. They included an ‘‘equal distance’’ model in their comparison, with $d_{a_1, a_2} = I(a_1 \neq a_2)$. In this article, the equal distance

model is referred to as ‘‘GYE,’’ and the Miyata, Miyazawa, and Yasunaga model as ‘‘GYM.’’

Quantities of particular interest, once the parameters of a codon model have been estimated, are the expected rates of synonymous (ρ_s) and nonsynonymous (ρ_a) substitution, and their ratio, ρ_s/ρ_a . Estimates of these expected rates can be derived from the estimated rate matrix, using the method described above (see equation B.2). They can be used to estimate the ratio K_a/K_s (or its inverse, K_s/K_a), as described by Goldman and Yang (1994). Note that the convention with codon models has been to scale \mathbf{Q} such that the expected number of substitutions per *codon* is one, rather than the expected number per *site* (as we propose above). Our convention allows parameter estimates from models of different orders to be compared more easily.

The Extension of Felsenstein’s Algorithm

The quantity in the denominator of equation 5, $\sum_{\tilde{x}_{\bullet, i}} P(x_{\bullet, i-N+1}, \dots, x_{\bullet, i-1}, \tilde{x}_{\bullet, i} | \Psi)$, is computed using the missing-data principle described in Appendix A. Let $Z_{u, i} = (X_{u, i-N+1}, \dots, X_{u, i})$ and $z_{u, i} = (x_{u, i-N+1}, \dots, x_{u, i})$. In addition, let $y_{u, i}$ be the set of observed characters at the leaves beneath node u for sites $i - N + 1, \dots, i - 1$ (that is, excluding site i), so that $P(y_{u, i} | Z_{u, i} = a)$ is the probability of those characters given that node u is assigned tuple a (compare to $x_{u, i}$, Appendix A). Now, let us say that tuple $a \in \Sigma^N$ *partially matches* tuple $b \in \Sigma^N$ if a and b share their first $N - 1$ characters, and let us denote this relationship as $a \approx b$. The sum in question can be computed using a version of Felsenstein’s algorithm (equation A.1) in which the base case of the recursion is altered as follows:

$$P(y_{u, i} | Z_{u, i} = a) = \begin{cases} I(z_{u, i} \approx a) & \text{if } u \in \mathcal{L} \\ \sum_{b \in \Sigma^N} P(b | a, \beta_v) P(y_{v, i} | Z_{v, i} = b) \\ \quad \times \sum_{c \in \Sigma^N} P(c | a, \beta_w) P(y_{w, i} | Z_{w, i} = c) & \text{if } u \in \mathcal{I}. \end{cases} \quad (\text{B.4})$$

As usual, a final step must be performed at the root r of the tree to obtain the desired value: $\sum_{\tilde{x}_{\bullet, i}} P(x_{\bullet, i-N+1}, \dots, x_{\bullet, i-1}, \tilde{x}_{\bullet, i} | \Psi) = \sum_a \pi_a P(y_{r, i} | Z_{r, i} = a)$.

Appendix C: Parameter Estimation by EM Derivation of the Update Rule

By the standard EM algorithm, each new estimate $(\hat{\mathbf{Q}}^{t+1}, \hat{\boldsymbol{\beta}}^{t+1})$ is determined from the previous estimate $(\hat{\mathbf{Q}}^t, \hat{\boldsymbol{\beta}}^t)$ as

$$(\hat{\mathbf{Q}}^{t+1}, \hat{\boldsymbol{\beta}}^{t+1}) = \arg \max_{\mathbf{Q}, \boldsymbol{\beta}} \sum_{z_{\circ}} P(z_{\circ} | z_{\bullet}, \hat{\mathbf{Q}}^t, \boldsymbol{\pi}, \boldsymbol{\tau}, \hat{\boldsymbol{\beta}}^t) \times \log P(z_{\bullet}, z_{\circ} | \mathbf{Q}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\beta}).$$

Taking advantage of independence of column tuples, the fact that each $P(z_{\bullet, j}, z_{\circ, j} | \mathbf{Q}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\beta})$ (for $1 \leq j \leq L'$) is simply a product over the edges of the tree, and the fact

that π and τ are constants in the maximization, this equation can be rewritten as

$$\begin{aligned} (\hat{\mathbf{Q}}^{t+1}, \hat{\boldsymbol{\beta}}^{t+1}) &= \arg \max_{\mathbf{Q}, \boldsymbol{\beta}} \sum_{z_o} P(z_o | z_{\bullet}, \hat{\mathbf{Q}}^t, \pi, \tau, \hat{\boldsymbol{\beta}}^t) \\ &\quad \times \log \left[\prod_{j=1}^{L'} \pi_{z_r, j} \prod_{u \in \mathcal{V} - \{r\}} P(z_{u, j} | z_{\sigma(u), j}, \beta_u) \right] \\ &= \arg \max_{\mathbf{Q}, \boldsymbol{\beta}} \sum_{z_o} P(z_o | z_{\bullet}, \hat{\mathbf{Q}}^t, \pi, \tau, \hat{\boldsymbol{\beta}}^t) \\ &\quad \times \sum_{j=1}^{L'} \sum_{u \in \mathcal{V} - \{r\}} \log P(z_{u, j} | z_{\sigma(u), j}, \beta_u) \quad (\text{C.1}) \end{aligned}$$

Now, with $\hat{\boldsymbol{\Psi}}^t = (\hat{\mathbf{Q}}^t, \pi, \tau, \hat{\boldsymbol{\beta}}^t)$ for notational convenience, let $E[S(b, a, u) | z_{\bullet}, z_o, \hat{\boldsymbol{\Psi}}^t]$ be the expected number of substitutions of $b \in \Sigma^N$ for $a \in \Sigma^N$ on the edge above node u , given z_{\bullet}, z_o , and the previous parameter estimates, and let

$$\begin{aligned} E[S(b, a, u) | z_{\bullet}, \hat{\boldsymbol{\Psi}}^t] &= \sum_{z_o} E[S(b, a, u) | z_{\bullet}, z_o, \hat{\boldsymbol{\Psi}}^t] \\ &\quad \times P(z_o | z_{\bullet}, \hat{\boldsymbol{\Psi}}^t) \quad (\text{C.2}) \end{aligned}$$

be the expectation of the same quantity considering all possible values of the latent variables. Because $z_{u, j}, z_{\sigma(u), j} \in \Sigma^N$, the terms in equation C.1 of the form $\log P(z_{u, j} | z_{\sigma(u), j}, \beta_u)$ can be grouped by pairs $a, b \in \Sigma^N$ such that $z_{u, j} = b$ and $z_{\sigma(u), j} = a$. Using this idea, and substituting from equation C.2, we obtain equation 2 from the text:

$$\begin{aligned} (\hat{\mathbf{Q}}^{t+1}, \hat{\boldsymbol{\beta}}^{t+1}) &= \arg \max_{\mathbf{Q}, \boldsymbol{\beta}} \sum_{z_o} P(z_o | z_{\bullet}, \hat{\boldsymbol{\Psi}}^t) \\ &\quad \times \sum_{u \in \mathcal{V} - \{r\}} \sum_{a, b \in \Sigma^N} E[S(b, a, u) | z_{\bullet}, z_o, \hat{\boldsymbol{\Psi}}^t] \log P(b | a, \beta_u) \\ &= \arg \max_{\mathbf{Q}, \boldsymbol{\beta}} \sum_{u \in \mathcal{V} - \{r\}} \sum_{a, b \in \Sigma^N} E[S(b, a, u) | z_{\bullet}, \hat{\boldsymbol{\Psi}}^t] \log P(b | a, \beta_u) \\ &= \arg \max_{\mathbf{Q}, \boldsymbol{\beta}} \sum_{u \in \mathcal{V} - \{r\}} \sum_{a, b \in \Sigma^N} E[S(b, a, u) | z_{\bullet}, \hat{\mathbf{Q}}^t, \pi, \tau, \hat{\boldsymbol{\beta}}^t] \\ &\quad \times \log([\exp(\mathbf{Q}\boldsymbol{\beta}_u)]_{a, b}) \quad (2) \end{aligned}$$

The Expectation Step

The expectation (E) step of the EM algorithm consists of computing the expected numbers of substitutions of each type along each edge of the tree (equation C.2). It is not hard to show that, when independence of column tuples is assumed, these values can be obtained directly from the posterior probabilities of each type of substitution at each tuple of sites:

$$E[S(b, a, u) | z_{\bullet}, \hat{\boldsymbol{\Psi}}^t] = \sum_{j=1}^{L'} P(Z_{u, j} = b, Z_{\sigma(u), j} = a | z_{\bullet, j}, \hat{\boldsymbol{\Psi}}^t). \quad (\text{C.3})$$

These posterior probabilities can be computed efficiently by combining the intermediate values of Felsen-

stein's algorithm with a set of complementary values, in an "inside-outside" procedure analogous to the "forward-backward" algorithm used with hidden Markov models (Durbin et al. 1998). (Both are instances of a general "sum-product" algorithm for graphical models [Jordan 2004]). Let $z_{\bar{u}, j}$ denote the observed characters at site j *not* beneath node u ; that is, $z_{\bar{u}, j} = z_{r, j} - z_{u, j}$. The posterior probability of base a at site j is

$$P(Z_{u, j} = a | z_{\bullet, j}, \hat{\boldsymbol{\Psi}}^t) = \frac{P(z_{\bar{u}, j} | Z_{u, j} = a)P(z_{\bar{u}, j}, Z_{u, j} = a)}{P(z_{\bullet, j} | \hat{\boldsymbol{\Psi}}^t)}, \quad (\text{C.4})$$

where the probabilities in the numerator are implicitly conditioned on $\hat{\boldsymbol{\Psi}}^t$. The quantities of the form $P(z_{\bar{u}, j}, Z_{u, j} = a)$ can be computed recursively, as follows. Let v be the parent of u and let w be the sibling of u ($\sigma(u) = \sigma(w) = v$). Then,

$$\begin{aligned} P(z_{\bar{u}, j}, Z_{u, j} = a) &= \begin{cases} \pi_a & \text{if } u = r \\ \sum_{b, c} P(z_{\bar{v}, j}, Z_{v, j} = b)P(a | b, \beta_u) \\ \quad \times P(z_{w, j} | Z_{w, j} = c)P(c | b, \beta_w) & \text{otherwise.} \end{cases} \quad (\text{C.5}) \end{aligned}$$

Equation C.4 yields quantities equivalent to the marginal probabilities for ancestral states that were introduced by Yang, Kumar, and Nei (1995). The algorithm of Koshi and Goldstein (1996) for computing these probabilities appears to be essentially equivalent to the inside-outside procedure described here.

The probability of each type of substitution along each edge is given by

$$\begin{aligned} P(Z_{u, j} = b, Z_{\sigma(u), j} = a | z_{\bullet, j}, \hat{\boldsymbol{\Psi}}^t) &= \frac{P(z_{\bar{u}, j} | Z_{u, j} = b)P(b | a, \beta_u)}{\sum_c P(z_{\bar{u}, j} | Z_{u, j} = c)P(c | a, \beta_u)} \times P(Z_{\sigma(u), j} = a | z_{\bullet, j}, \hat{\boldsymbol{\Psi}}^t). \quad (\text{C.6}) \end{aligned}$$

See Friedman et al. (2002) for an alternative derivation of these posterior probabilities, which assumes reversibility.

When rate variation is allowed, separate expected numbers of substitutions are required for each rate category l (equation 3). In this case,

$$\begin{aligned} E[S(b, a, u, l) | z_{\bullet}, \hat{\boldsymbol{\Psi}}^t, \hat{\boldsymbol{\alpha}}^t] &= \sum_{j=1}^{L'} P(Z_{u, j} = b, Z_{\sigma(u), j} = a | Y_j = l, z_{\bullet, j}, \hat{\boldsymbol{\Psi}}^t, \hat{\boldsymbol{\alpha}}^t) \\ &\quad \times P(Y_j = l | z_{\bullet, j}, \hat{\boldsymbol{\Psi}}^t, \hat{\boldsymbol{\alpha}}^t) \quad (\text{C.7}) \end{aligned}$$

where Y_j is a random variable indicating the rate category of the j th tuple and

$$P(Y_j = l | z_{\bullet, j}, \hat{\boldsymbol{\Psi}}^t, \hat{\boldsymbol{\alpha}}^t) = \frac{P(z_{\bullet, j}, Y_j = l | \hat{\boldsymbol{\Psi}}^t, \hat{\boldsymbol{\alpha}}^t)}{P(z_{\bullet, j} | \hat{\boldsymbol{\Psi}}^t, \hat{\boldsymbol{\alpha}}^t)}$$

is the posterior probability of $Y_j = l$. In addition, $P(Z_{u, j} = b, Z_{\sigma(u), j} = a | Y_j = l, z_{\bullet, j}, \hat{\boldsymbol{\Psi}}^t, \hat{\boldsymbol{\alpha}}^t)$ is simply $P(Z_{u, j} = b, Z_{\sigma(u), j} = a | z_{\bullet, j}, \hat{\boldsymbol{\Psi}}^t)$, where $\hat{\boldsymbol{\Psi}}^t = (\hat{\mathbf{Q}}^t, \pi, \tau, r\hat{\boldsymbol{\beta}}^t)$. This probability can

be obtained as above (equations C.4, C.5, and C.6) but with $\hat{\Psi}'_l$ in place of $\hat{\Psi}'$.

The Maximization Step

The maximization (M) step consists of finding the parameter values $(\hat{\mathbf{Q}}, \hat{\boldsymbol{\beta}})$ that maximize the function

$$f(\mathbf{Q}, \boldsymbol{\beta}) = \sum_{u \in \mathcal{V} - \{r\}} \sum_{a, b \in \Sigma^N} C_{b, a, u} \log([\exp(\mathbf{Q}\boldsymbol{\beta}_u)]_{a, b}), \quad (\text{C.8})$$

where the terms of the form $C_{b, a, u} = E[S(b, a, u) | z_{\bullet}, \hat{\mathbf{Q}}^t, \boldsymbol{\pi}, \boldsymbol{\tau}, \hat{\boldsymbol{\beta}}^t]$ can be considered constants (see equation 2). This can be accomplished with a quasi-Newton algorithm, given the partial derivatives of $f(\mathbf{Q}, \boldsymbol{\beta})$ with respect to the free parameters of the model. The partial derivative with respect to a free parameter \mathcal{P} is

$$\frac{\partial f(\mathbf{Q}, \boldsymbol{\beta})}{\partial \mathcal{P}} = \sum_{u \in \mathcal{V} - \{r\}} \sum_{a, b \in \Sigma^N} C_{b, a, u} \frac{\partial [\exp(\mathbf{Q}\boldsymbol{\beta}_u)]_{a, b}}{[\exp(\mathbf{Q}\boldsymbol{\beta}_u)]_{a, b}}. \quad (\text{C.9})$$

Assume \mathbf{Q} is diagonalizable as $\mathbf{S}\boldsymbol{\Lambda}\mathbf{S}^{-1}$, and let the elements of \mathbf{S} and \mathbf{S}^{-1} be denoted $\{s_{a, b}\}$ and $\{s'_{a, b}\}$, respectively, with $a, b \in \Sigma^N$. Similarly, let the diagonal elements of $\boldsymbol{\Lambda}$ be denoted $\{\lambda_a\}$, $a \in \Sigma^N$. (Note that elements of \mathbf{S} , \mathbf{S}^{-1} , and $\boldsymbol{\Lambda}$ may be complex-valued in the case of unrestricted models.) Thus, $[\exp(\mathbf{Q}\boldsymbol{\beta}_u)]_{a, b} = \sum_{c \in \Sigma^N} s_{a, c} e^{\lambda_c \beta_u} s'_{c, b}$, and if \mathcal{P} is a branch-length parameter β_u ,

$$\begin{aligned} \frac{\partial f(\mathbf{Q}, \boldsymbol{\beta})}{\partial \beta_u} &= \sum_{a, b \in \Sigma^N} C_{b, a, u} \frac{\frac{\partial}{\partial \beta_u} \sum_{c \in \Sigma^N} s_{a, c} e^{\lambda_c \beta_u} s'_{c, b}}{[\exp(\mathbf{Q}\boldsymbol{\beta}_u)]_{a, b}} \\ &= \sum_{a, b \in \Sigma^N} C_{b, a, u} \frac{\sum_{c \in \Sigma^N} s_{a, c} \lambda_c e^{\lambda_c \beta_u} s'_{c, b}}{[\exp(\mathbf{Q}\boldsymbol{\beta}_u)]_{a, b}}. \end{aligned} \quad (\text{C.10})$$

If \mathcal{P} is a rate-matrix parameter, then the partial derivatives of the form $\frac{\partial}{\partial \mathcal{P}} [\exp(\mathbf{Q}\boldsymbol{\beta}_u)]_{a, b}$ are the elements of the matrix $\frac{\partial}{\partial \mathcal{P}} \exp(\mathbf{Q}\boldsymbol{\beta}_u)$, which Schadt and Lange (2002) have shown, using a derivation based on the Cauchy integral formula, to be given by

$$\frac{\partial}{\partial \mathcal{P}} \exp(\mathbf{Q}\boldsymbol{\beta}_u) = \mathbf{S} \left[\mathbf{F} \circ \left(\mathbf{S}^{-1} \frac{\partial}{\partial \mathcal{P}} (\mathbf{Q}\boldsymbol{\beta}_u) \mathbf{S} \right) \right] \mathbf{S}^{-1} \quad (\text{C.11})$$

where \circ denotes the Hadamard (pointwise) product of two matrices, and $\mathbf{F} = \{f_{a, b}\}$ is defined as

$$f_{a, b} = \begin{cases} \beta_u \exp(\lambda_a \beta_u) & \text{if } \lambda_a = \lambda_b \\ \frac{\exp(\lambda_a \beta_u) - \exp(\lambda_b \beta_u)}{\lambda_a - \lambda_b} & \text{otherwise.} \end{cases}$$

The partial derivative for each branch-length parameter can be obtained in $O(d^{3N})$ time, but the derivative for each rate-matrix parameter requires $O(nd^{3N})$ time. With context-dependent models, the computation of rate-matrix derivatives is a bottleneck for the EM algorithm, and it is worthwhile to approximate them. Using only the first few

terms of the Taylor expansion of $\exp(\mathbf{Q}t)$ turns out to accelerate the computation considerably. For example, if the first four terms are used,

$$\begin{aligned} \frac{\partial}{\partial \mathcal{P}} \exp(\mathbf{Q}t) &\approx \frac{\partial}{\partial \mathcal{P}} \left(\mathbf{I} + \mathbf{Q}t + \frac{\mathbf{Q}^2 t^2}{2} + \frac{\mathbf{Q}^3 t^3}{6} \right) \\ &= \mathbf{Q}'t + \frac{t^2}{2} (\mathbf{Q}'\mathbf{Q} + \mathbf{Q}\mathbf{Q}') \\ &\quad + \frac{t^3}{6} (\mathbf{Q}'\mathbf{Q}^2 + \mathbf{Q}\mathbf{Q}'\mathbf{Q} + \mathbf{Q}^2\mathbf{Q}'). \end{aligned} \quad (\text{C.12})$$

where $\mathbf{Q}' = \frac{\partial}{\partial \mathcal{P}} \mathbf{Q}$. Because \mathbf{Q}' is very sparse for our models (it has between 2 and 8 non-zero elements, if the constraint on scaling is temporarily relaxed), matrix multiplications of the form $\mathbf{X}\mathbf{Q}'\mathbf{Y}$ can be accomplished in $O(d^{2N})$ time. In addition, the matrices \mathbf{Q}^k ($k \geq 2$) can be precomputed and reused for all rate-matrix parameters. As a result, the time required to compute each rate-matrix derivative is effectively reduced by a factor of d^N (provided the number of rate-matrix parameters is large enough, as in our case).

Unfortunately, even when more than four terms of the Taylor expansion are used, the approximate derivatives are not sufficient for precise maximization of $f(\mathbf{Q}, \boldsymbol{\beta})$. They are useful, however, in early iterations of the EM algorithm, in which approximate maximization is adequate. Our software switches to exact derivatives when the approximate ones no longer lead to significant improvements in likelihood.

When rate variation is allowed, $f(\mathbf{Q}, \boldsymbol{\beta})$ becomes

$$f(\mathbf{Q}, \boldsymbol{\beta}) = \sum_{l=1}^k \sum_{u \in \mathcal{V} - \{r\}} \sum_{a, b \in \Sigma^N} C_{b, a, u, l} \log([\exp(\mathbf{Q}_l \boldsymbol{\beta}_u)]_{a, b}) \quad (\text{C.13})$$

where $C_{b, a, u, l} = E[S(b, a, u, l) | z_{\bullet}, \hat{\mathbf{Q}}^t, \boldsymbol{\pi}, \boldsymbol{\tau}, \hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\alpha}}^t]$ (equation 3). The computation of branch-length and rate-matrix derivatives remains essentially unchanged, but now involves an additional sum over the k rate categories, resulting in a factor of k increase in computational time. In addition, the partial derivative with respect to the shape parameter α is now required. It can be computed as

$$\begin{aligned} \frac{\partial f(\mathbf{Q}, \boldsymbol{\beta})}{\partial \alpha} &= \sum_{l=1}^k \sum_{u \in \mathcal{V} - \{r\}} \sum_{a, b \in \Sigma^N} C_{b, a, u, l} \\ &\quad \times \frac{\sum_{c \in \Sigma^N} s_{a, c} \lambda_c \beta_u e^{\lambda_c \beta_u} s'_{c, b} \frac{\partial r_l}{\partial \alpha}}{[\exp(\mathbf{Q}_l \boldsymbol{\beta}_u)]_{a, b}}. \end{aligned} \quad (\text{C.14})$$

We simply obtain $\frac{\partial r_l}{\partial \alpha}$ numerically.

Performance

The performance of the EM algorithm depends on many factors, including the size of the alignment (number of species n and number of sites L), the order of the model, the number of rate-matrix parameters, and the suitability of the starting values for parameters. Nevertheless, a rough idea of its performance can be obtained from the running times recorded in our experiments with the AR data set ($n = 9$ species, $L = 162,743$ sites). Experiments were performed on a desktop system with a 2.4 GHz Pentium

IV processor (500 MB of RAM, Red Hat Linux 7.3), using an implementation of the algorithm written in C and compiled with gcc. The software relies on the LAPACK package (Anderson et al. 1999) for matrix diagonalization, and uses our own implementation of the BFGS quasi-Newton algorithm, based on the one described by Press et al. (1992). Parameters were initialized to arbitrary values (0.1 for all branch-length parameters, rate matrix parameters such that transitions had rates five times greater than transversions). Without rate variation, first-order models required 48.5 s (HKY), 49.2 s (REV), and 83.4 s (UNR), second-order models took between 12.3 min (R2S) and 47.4 min (U2), and third-order models between 18.9 h (R3S) and 6.6 days (U3). (Unrestricted models took considerably longer than reversible ones, in part because of complex arithmetic in derivative computations.) For comparison, the baseml program of the PAML package (version 3.13; Yang 1997) required 95 seconds for HKY and 133 seconds for REV, using the same data set and hardware. When a quasi-Newton algorithm was used in place of EM for context-dependent models, running times increased substantially, e.g., from 39.5 min to 18.0 h with the U2S model. The performance of the EM algorithm was significantly degraded when rate variation was allowed ($k = 4$ rate categories), and it depended heavily on the initial value of α ; e.g., with a starting value of $\alpha = 1$, the REV model took 672 seconds, but with a starting value of $\alpha = 8$, it required only 158 seconds. We did not try to fit context-dependent models using poor starting values of α ; with a starting value of 10, the U2S model took 93.4 min and the U3S model took 8.2 days (compared to 39.5 min and 4.8 days, respectively, without rate variation).

The EM algorithm generally required between 20 and 60 iterations to converge (sometimes more with rate variation). The M step was essentially instantaneous for first-order models, and for second-order models (without rate variation), it typically required between 1 and 5 s with the approximate derivatives, and between 5 and 30 s without them (occasionally 100–200 s, as when first switching to exact derivatives). The time of the M step varied widely for third-order models, depending on the number of iterations required to reach convergence; typical times with the approximate derivatives were between 1 and 20 min, and typical times without them were 20–100 min, but sometimes 8–10 h were required, and occasionally, 48 h or more. The E step, in contrast, was very consistent from one iteration to the next; it required 1.9 s for the UNR model, 38 s for the U2S model, and 27 min for the U3S model (without rate variation). Allowing rate variation increased the time of the E step by a factor of k , and the cost of the M step by at least a factor of k (usually somewhat more). For this reason, our software first fits a model approximately without rate variation, then switches to the discrete gamma model.

It is worth noting that the most computationally demanding parts of the algorithm—the computation of posterior probabilities of substitution in the E step, and of partial derivatives with respect to free parameters in the M step—are readily parallelizable. We expect that a parallel implementation of the algorithm would considerably reduce the time required to fit third-order models in particular.

Acknowledgments

We thank Eric Green and the NIH Intramural Sequencing Center for permission to use the CFTR sequence data prior to publication; Webb Miller for preparing the multiple alignment for the CFTR data set; Jim Kent, Matt Schwartz, and the UC Santa Cruz genome browser team for advice on retrieving and scoring the mRNA sequences; David Heckerman for a helpful discussion regarding Markov dependence of sites; and Mathieu Blanchette for comments on the manuscript. Nick Goldman encouraged us to write this article in the first place, and he provided several useful references when we finally began to put it together. Ziheng Yang's PAML package was helpful in testing the correctness of our software, and we borrowed its subroutines for computing discrete gamma mean rates. Finally, we thank the anonymous reviewers for several very good suggestions, particularly regarding model testing. Funding was provided by the Howard Hughes Medical Institute (D.H.) and National Human Genome Research Institute grant IP41HG02371 (A.S.).

Literature Cited

- Adachi, J., and M. Hasegawa. 1996. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. The Institute of Statistical Mathematics, Tokyo.
- Anderson, E., Z. Bai, C. Bischoff, et al. (11 co-authors). 1999. LAPACK Users' Guide, 3rd edition. Society for Industrial and Applied Mathematics, Philadelphia, Penna.
- Arndt, P. F., C. B. Burge, and T. Hwa. 2002. DNA sequence evolution with neighbor-dependent mutation. Pp. 32–38 in Myers, G., S. Hannenhalli, S. Istrail, P. Pevzner, and M. Waterman, eds. Proceedings of the Sixth Annual International Conference on Computational Biology. Association for Computing Machinery, New York.
- Bernardi, G. 2000. The compositional evolution of vertebrate genomes. *Gene* 259:31–43.
- Blake, R. D., S. T. Hess, and J. Nicholson-Tuell. 1992. The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *J. Mol. Evol.* 34:189–200.
- Blanchette, M., W. J. Kent, C. Riener, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* (in press).
- Bray, N., and L. Pachter. 2003. MAVID multiple alignment server. *Nucleic Acids Res.* 31:3525–3526.
- Bulmer, M. 1986. Neighboring base effects on substitution rates in pseudo genes. *Mol. Biol. Evol.* 3:322–329.
- Cooper, G. M., M. Brudno, NISC Comparative Sequencing Program, E. D. Green, S. Batzoglou, and A. Sidow. 2003. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* 13:813–820.
- Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. A model of evolutionary change in proteins. Pp. 345–352 in Atlas of protein sequence and structure, vol. 5. National Biomedical Research Foundation, Washington, D.C.
- Dempster, A., N. Laird, and D. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. Ser. B*, 39:1–38.
- Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge, U.K.

- Ehrlich, M., X.-Y. Zhang, and N. M. Inamdar. 1990. Spontaneous deamination of cytosine and 5-methylcytosine residues in DNA and replacement of 5-methylcytosine residues with cytosine residues. *Mutat. Res.* **238**:277–286.
- Felsenstein, J. 1981. Evolutionary trees from DNA sequences. *J. Mol. Evol.* **17**:368–376.
- . 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle. Available from <http://evolution.genetics.washington.edu/phylip.html>.
- Felsenstein, J., and G. A. Churchill. 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13**:93–104.
- Friedman, N., M. Ninio, I. Pe'er, and T. Pupko. 2002. A structural EM algorithm for phylogenetic inference. *J. Comp. Biol.* **9**:331–353.
- Fryxell, K. J., and E. Zuckerkandl. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.* **17**:1371–1383.
- Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**:725–735.
- Goldman, N., J. L. Thorne, and D. T. Jones. 1996. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol.* **263**:196–208.
- Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. *Science* **185**:862–864.
- Green, P., B. Ewing, W. Miller, P. J. Thomas, NISC Comparative Sequencing Program, and E. D. Green. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* **33**:514–517.
- Hardison, R., K. M. Roskin, S. Yang, et al. 2003. Covariation in frequencies of substitution, deletion, transposition and recombination during eutherian evolution. *Genome Res.* **13**:13–26.
- Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**:160–174.
- Hastie, T., R. Tibshirani, and J. Friedman. 2001. *The Elements of Statistical Learning*. Springer-Verlag, New York.
- Hess, S. T., J. D. Blake, and R. D. Blake. 1994. Wide variations in neighbor-dependent substitution rates. *J. Mol. Biol.* **236**:1022–1033.
- Huelsenbeck, J., and B. Rannala. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* **276**:227–232.
- Jensen, J. L., and A.-M. K. Pedersen. 2000. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv. Appl. Prob.* **32**:499–517.
- Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**:275–282.
- Jordan, M. I. (ed). 1999. *Learning in Graphical Models*. MIT Press, Cambridge, Mass.
- . 2004. Graphical models. *Stat. Sci.* (Special Issue on Bayesian Statistics). In press.
- Karlin, S., and H. M. Taylor. 1975. *A First Course in Stochastic Processes*, 2nd edition. Academic Press, San Diego, CA.
- Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. 2002. The human genome browser at UCSC. *Genome Res.* **12**:996–1006.
- Koshi, J. M., and R. M. Goldstein. 1996. Probabilistic reconstruction of ancestral protein sequences. *J. Mol. Evol.* **42**:313–320.
- Liò, P., and N. Goldman. 1998. Models of molecular evolution and phylogeny. *Genome Res.* **8**:1233–1244.
- Miyata, T., S. Miyazawa, and T. Yasunaga. 1979. Two types of amino acid substitution in protein evolution. *J. Mol. Evol.* **12**:219–236.
- Morton, B. R. 1997. The influence of neighboring base composition on substitutions in plant chloroplast coding sequences. *Mol. Biol. Evol.* **14**:189–194.
- Morton, B. R., and M. T. Clegg. 1995. Neighboring base composition is strongly correlated with base substitution bias in a region of the chloroplast genome. *J. Mol. Evol.* **41**:597–603.
- Morton, B. R., V. M. Oberholzer, and M. T. Clegg. 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., Y. Weiss, and M. I. Jordan. 1999. Loopy belief-propagation for approximate inference: an empirical study. Pp. 467–475 in K. B. Laskey, and H. Prade, eds. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI)*. Morgan Kaufmann, San Mateo, Calif.
- Muse, S. 1995. Evolutionary analyses of DNA sequences subject to constraints on secondary structure. *Genetics* **139**:1429–1439.
- Muse, S. V., and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **11**:715–724.
- Neyman, J. 1971. Molecular studies of evolution: a source of novel statistical problems. Pp. 1–27 in S. Gupta, and J. Yackel, eds. *Statistical decision theory and related topics*, Academic Press, New York.
- Notredame, C., D. Higgins, and J. Heringa. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**:205–217.
- Olsen, G. J., H. Matsuda, R. Hagstrom, and R. Overbeek. 1994. fastDNAm1: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.* **10**:41–48.
- Pedersen, A.-M. K., and J. L. Jensen. 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., C. Wiuf, and F. B. Christiansen. 1998. A codon-based model designed to describe lentiviral evolution. *Mol. Biol. Evol.* **15**:1069–1081.
- Pedersen, J. S., and J. Hein. 2003. Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics* **19**:219–227.
- Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. 1992. *Numerical Recipes in C: The art of scientific computing*, 2nd edition. Cambridge University Press, Cambridge, U.K.
- Pruitt, K. D., and D. R. Maglott. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**:137–140.
- Rzhetsky, A. 1995. Estimating substitution rates in ribosomal RNA genes. *Genetics* **141**:771–783.
- Schadt, E., and K. Lange. 2002. Codon and rate variation models in molecular phylogeny. *Mol. Biol. Evol.* **19**:1534–1549.
- Schöniger, M., and A. von Haeseler. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phyl. Evol.* **3**:240–247.
- Schöniger, M., and A. von Haeseler. 1995. Performance of the maximum likelihood, Neighbor Joining, and maximum

- parsimony methods when sequence sites are not independent. *Syst. Biol.* **44**:533–547.
- Schwartz, G. 1979. Estimating the dimension of a model. *Ann. Stat.* **6**:461–464.
- Siepel, A., and D. Haussler. 2004. Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comp. Biol.* (in press).
- Swofford, D. L. 2002. PAUP*: phylogenetic analysis using parsimony (and other methods), version 4. Sinauer Associates, Sunderland, Mass.
- Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* **17**:57–86.
- Thomas, J. W., J. W. Touchman, R. W. Blakesley et al. (11 co-authors). 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**:788–793.
- Tillier, E. R. M., and R. A. Collins. 1995. Neighbor-Joining and maximum likelihood with RNA sequences: addressing the interdependence of sites. *Mol. Biol. Evol.* **12**:7–15.
- Wainwright, M., T. Jaakkola, and A. Willsky. 2001. Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Transactions on Information Theory* **49**(5):1120–1146.
- Whelan, S., P. Liò, and N. Goldman. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet.* **17**:262–272.
- Yang, Y.-W., Y. Chen, and W.-H. Li. 2002. The influence of adjacent nucleotides on the pattern of nucleotide substitution in mitochondrial introns of angiosperms. *J. Mol. Evol.* **55**:111–115.
- Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**:1396–1401.
- . 1994a. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**:105–111.
- . 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**:306–314.
- . 1995. A space-time process model for the evolution of DNA sequences. *Genetics* **139**:993–1005.
- . 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* **13**:555–556.
- Yang, Z., and S. Kumar. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. *Mol. Biol. Evol.* **13**:650–659.
- Yang, Z., and D. Roberts. 1995. On the use of nucleic acid sequences to infer early branching in the tree of life. *Mol. Biol. Evol.* **12**:451–458.
- Yang, Z., N. Goldman, and A. Friday. 1994. Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11**:316–224.
- Yang, Z., S. Kumar, and M. Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**:1641–1650.
- Yang, Z., R. Nielsen, and M. Hasegawa. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.* **15**:1600–1611.
- Yang, Z., R. Nielsen, N. Goldman, and A.-M. K. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**:431–449.
- Yedidia, J., W. Freeman, and Y. Weiss. 2000. Generalized belief propagation. *Advances in Neural Processing Systems* **13**:689–695.
- Zhao, Z., and E. Boerwinkle. 2002. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res.* **12**:1679–1686.

Hervé Phillippe, Associate Editor

Accepted October 10, 2003