A New Method for Characterizing Replacement Rate Variation in Molecular Sequences: Application of the Fourier and Wavelet Models to Drosophila and Mammalian Proteins

Pavel Morozov,*,[1] Tatyana Sitnikova,[†,1] Gary Churchill,[‡] Francisco José Ayala[§] and Andrey Rzhetsky**

**Columbia Genome Center, Columbia University, New York, New York 10032, [†]Eisai Research Institute, GEFA Biology Group, Boston, Massachusetts 02138, [‡]The Jackson Laboratory, Bar Harbor, Maine 04609, [§]New York, New York 10013 and **Department of Medical Informatics, Columbia University, New York, New York 10032*

Manuscript received March 27, 1998
Accepted for publication September 22, 1999

ABSTRACT

We propose models for describing replacement rate variation in genes and proteins, in which the profile of relative replacement rates along the length of a given sequence is defined as a function of the site number. We consider here two types of functions, one derived from the cosine Fourier series, and the other from discrete wavelet transforms. The number of parameters used for characterizing the substitution rates along the sequences can be flexibly changed and in their most parameter-rich versions, both Fourier and wavelet models become equivalent to the unrestricted-rates model, in which each site of a sequence alignment evolves at a unique rate. When applied to a few real data sets, the new models appeared to fit data better than the discrete gamma model when compared with the Akaike information criterion and the likelihood-ratio test, although the parametric bootstrap version of the Cox test performed for one of the data sets indicated that the difference in likelihoods between the two models is not significant. The new models are applicable to testing biological hypotheses such as the statistical identity of rate variation profiles among homologous protein families. These models are also useful for determining regions in genes and proteins that evolve significantly faster or slower than the sequence average. We illustrate the application of the new method by analyzing human immunoglobulin and Drosophilid alcohol dehydrogenase sequences.

WHILE variation in the rate of amino acid replacement across sites within protein sequences has been frequently observed, mathematical modeling of this phenomenon remains a challenging problem. The difficulty lies mainly in choosing the right trade-off between the parameter richness of the model, quality of statistical inference (more parameter-rich models tend to provide parameter estimates with larger variance), and the computational cost of applying the model to real data, the optimum balance being usually different for different data sets and available computational facilities. Earlier models of rate variation had few parameters, while the number of model parameters steadily increased in the more recent models (*e.g.*, see Fitch and Margoliash 1967; Fitch and Markowitz 1970; Golding 1983; Jin and Nei 1990; Takahata 1991; Felsenstein 1993; Felsenstein and Churchill 1996; Wakeley 1993; Yang 1993; Yang and Wang 1995; Kelly and Rice 1996; Lake 1998). While in most of these formulations the number of rate variation parameters is fixed and relatively small, they have proven useful in a number of biological applications. However, we believe that due to steady accumulation of sequence data and stable improvement of computation facilities, more parameter-rich models are likely to eventually become advantageous.

Here we introduce a model that describes rate variation along sites as a function of each individual site. The number of parameters is not predetermined, but rather selected separately for each data set, ranging from one to the total number of sites in the sequence minus one. We introduce two model variations: wavelet and Fourier. We illustrate their application to human immunoglobulin and Drosophilid alcohol dehydrogenase sequences.

Maximum-likelihood calculations assuming unequal rate of substitution across sites: The long-term evolution of an individual site within a protein sequence can be conveniently modeled as a Markov chain in continuous time (*e.g.*, Kendall 1956). Label the 20 amino acids by integers 1, 2, . . . , 20 and denote by $X(t)$ the label of amino acid occupying a fixed site in the protein at time t ($t \geq 0$). The Markov process $X(t)$ thus has integer outcomes between 1 and 20, and there are well-devel-

Corresponding author: Andrey Rzhetsky, Columbia Genome Center, Columbia University, 1150 St. Nicholas Ave., Unit 109, New York, NY 10032. E-mail: andrey@genome2.cpmc.columbia.edu

[1] These authors contributed equally to this work.

oped techniques for computing the probability of replacing of amino acid i with amino acid j after time t, $P_{ij}(t)$, which is defined, assuming a time-homogeneous process, as

$$\text{Prob}\ \{X(t+s) = j \mid X(s) = i\} = P_{ij}(t) \quad \text{for all } s, t \geq 0; \quad i, j \in \{1, 2, \ldots, 20\}, \quad (1)$$

where $\{P_{ij}(t)\} = \mathbf{P}(t)$ is a matrix of transition probabilities.

Given a matrix of *instantaneous* transition probabilities, $\mathbf{Q}$, $\mathbf{P}(t)$ is computed as $\exp\{\mathbf{Q}t\}$. Assuming that our Markov process has an equilibrium state, we introduce a row vector $\boldsymbol{\pi}$ whose entries represent the equilibrium amino acid frequencies associated with $\mathbf{Q}$. By definition, the vector $\boldsymbol{\pi}$ can be obtained by solving the system of linear equations $\boldsymbol{\pi}\mathbf{Q} = 0$ and $\sum_{i=1}^{20} \pi_i = 1$.

$\mathbf{Q}$ and t will always appear in a likelihood function as a product and thus cannot be estimated separately. We therefore normalize $\mathbf{Q}$ (*e.g.*, see Yang and Wang 1995) such that

$$-\sum_{i=1}^{20} \pi_i Q_{ii} = 1 \text{ (by definition, } Q_{ii} = -\sum_{j \neq i} Q_{ij} \text{ for all } i). \quad (2)$$

Once $\mathbf{Q}$ is normalized, t will be measured in terms of the expected number of amino acid replacements per site. For example, for the simplest model of amino acid substitution, the Poisson model (Zuckerkandl and Pauling 1965), all off-diagonal entries of matrix $\mathbf{Q}$ are equal to $1/19$, and all diagonal entries of $\mathbf{Q}$ are equal to -1, while all entries in vector $\boldsymbol{\pi}$ are equal to $1/20$, while all off-diagonal elements in the matrix are different in the more advanced Jones *et al.* (1992) matrix. In this study we used only the Poisson (Zuckerkandl and Pauling 1965) and the Jones, Taylor, and Thornton (JTT; Jones *et al.* 1992) models, although the treatment of data under the Fourier/wavelet models would remain unchanged if other schemes of amino acid (Dayhoff *et al.* 1978; Kelly and Churchill 1996) or nucleotide substitution (*e.g.*, see Zharkikh 1994) were used. (We present here the mathematical treatment defined only for protein sequences, but it can be easily extended to nucleotide sequences by reducing the number of allowed characters to four and correspondingly changing the definition of matrix $\mathbf{Q}$.)

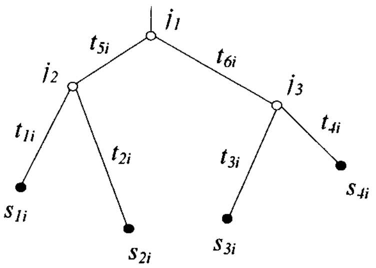

Figure 1.—A hypothetical four-sequence tree. (●) Pending vertices corresponding to the observed amino acids (s_{1i}, s_{2i}, s_{3i}, and s_{4i}) at the ith homologous site of present-day sequences 1, 2, 3, and 4. (○) Interior nodes corresponding to amino acids (j_1, j_2, and j_3) in the ith site of unknown ancestral sequences. The t_{ji}'s specify the expected length of the jth branch at the ith site of the protein alignment. $P_{ij}(t)$ in Equation 3 denotes the ijth entry of matrix $\mathbf{P}(t)$.

To illustrate the likelihood computation for a given tree, consider a hypothetical set $\mathbf{S}$ of four homologous contemporary protein sequences of length l aligned without gaps. We assume that the correct phylogenetic topology T for these sequences is known (see Figure 1). By definition, the likelihood L of observing the contemporary sequences under our model is Prob$\{\mathbf{S} \mid T, \boldsymbol{\theta}\}$ multiplied by an arbitrary positive constant (usually set to 1), where $\boldsymbol{\theta}$ is a set of specified values of model parameters. The probability of observing data at the ith site of the sequence alignment, given a set of parameter values $\boldsymbol{\theta}_i = \{\mathbf{Q}, t_{1i}, t_{2i}, \ldots, t_{6i}\}$ and topology T as shown in Figure 1, is computed as (Felsenstein 1981)

$$L_i = \sum_{j_1=1}^{20}\sum_{j_2=1}^{20}\sum_{j_3=1}^{20} \pi_{j_1} P_{j_1,j_2}(t_{5i}) P_{j_1,j_3}(t_{6i}) P_{j_2,s_{1i}}(t_{1i}) P_{j_2,s_{2i}}(t_{2i}) P_{j_3,s_{4i}}(t_{4i}), \quad (3)$$

where π_j is the frequency of the jth amino acid at the root of the tree, and the rest of the notations are as explained in the legend of Figure 1. The total likelihood L is a product of the probabilities of observing data at individual sites, and the maximum-likelihood estimates of the parameters are obtained by finding a set of parameter values that maximize the value of L.

Let us consider the topology shown in Figure 1 and assume that the set of branch lengths for the ith site, $\{t_{1i}, t_{2i}, \ldots, t_{6i}\}$, can be expressed as $\{t_1 c_i, t_2 c_i, \ldots, t_6 c_i\}$, where $\{t_1, t_2, \ldots, t_6\}$ is the set of expected branch lengths averaged over all l sites, and c_i is a nonnegative constant defining the *relative* replacement rate of the ith site. One can think of introducing $l - 1$ independent relative rate parameters $c_1, c_2, \ldots, c_{l-1}$ (one parameter for each of $l - 1$ sequence sites. Below we refer to this model as the *unrestricted rates model*; the relative rate of the lth site is not independent on other rates and is equal to $l - c_1 - \ldots - c_{l-1}$), although this is usually considered impractical because the number of free parameters appears large relative to the amount of data, thus making accurate estimation difficult. (There are other similarly formulated models. For example, Kelly and Rice (1996) assumed that c_i's are sampled identically and independently from the same distribution.)

There are several recognized ways to decrease the number of rate variation parameters (*e.g.*, see Fitch and Markowitz 1970; Fitch and Margoliash 1967; Golding 1983; Jin and Nei 1990; Takahata 1991; Felsenstein 1993; Felsenstein and Churchill 1996; Wakeley 1993; Yang 1993; Yang and Wang 1995; Lake 1998). One of the currently most popular approaches is to assume that the c_i's are independent and identically distributed random variables following a gamma distribution with the mean equal to 1 and the shape parameter estimated from the data (*e.g.*, see Golding 1983; Jin and Nei 1990; Yang 1993).

Here we suggest a different approach: we define c_i as a function of site number and a set of real-valued parameters $\{a_1, a_2, \ldots, a_k\}$, $c_i = f\{i; a_1, a_2, \ldots, a_{\mathrm{k}}\}$, where

$$f(i; a_1, a_2, \ldots, a_k)$$

$$= \left(1 + \sum_{j=1}^{k} a_j \psi\ (i, j)\right) \Big/ \left(1 + \sum_{n=1}^{k} \sum_{m=1}^{l} a_n \psi(n,m)/l\right). \quad (4)$$

The denominator of the expression is required to ensure that the average of $f(i; a_1, a_2, \ldots, a_k\}$ over $i = 1, 2, \ldots, l$ is always equal to 1. $\{\psi(i, j)\}_{j=1,k}$ are distinct basis functions, linear combination of which can precisely fit any relative substitution rate profile when $k = l - 1$; one can also get an approximation of the substitution rate profile when some of the least contributing basis functions are dropped. The function $f(i; a_1, a_2, \ldots, a_k\}$ cannot be observed directly, but rather has to be estimated using the maximum-likelihood approach.

Assuming that the potential readers of this article may not have had a prior exposure to the ideas of decomposition of functions, we first introduce relevant concepts in an intuitive way.

Basis vectors and functions, function decomposition: It might be easier to start by considering decomposition of a vector (Strung 1992). Every three-dimensional vector (x, y, z) can be decomposed into a sum of three "basis" vectors (1, 0, 0) (0, 1, 0), and (0, 0, 1). That is, (1, 0, 0) multiplied by x is $(x, 0, 0)$, (0, 1, 0) multiplied by y is $(0, y, 0)$, and (0, 0, 1) multiplied by z is $(0, 0, z)$. The sum is (x, y, z). The choice of the basis vectors is not unique: it is possible to find an infinite number of distinct sets of basis vectors. There are good reasons (convenience and simplicity) to use *orthogonal* sets of basis vectors, where vectors within each pair are perpendicular. Two vectors, (x_1, y_1, z_1) and (x_2, y_2, z_2), are called orthogonal or perpendicular if their inner product, defined as $x_1 x_2 + y_1 y_2 + z_1 z_2$, is equal to zero. It is easy to check that vectors (1, 0, 0) (0, 1, 0), and (0, 0, 1) are indeed pairwise orthogonal.

Decomposition of a vector into basis vectors is mathematically indistinguishable from decomposition of a discrete function into basis functions. We can simply view our three-dimensional vector (x, y, z), as a discrete function, $f(1) = x$, $f(2) = y$, and $f(3) = z$, and decompose this function into three basis functions, $\psi_1 = (1, 0, 0)$, $\psi_2 = (0, 1, 0)$, and $\psi_3 = (0, 0, 1)$: $f = x\ \psi_1 + y\ \psi_2 + z\ \psi_3$. Therefore, discrete basis functions are nothing but compact representations of basis vectors, and there are an infinite number of distinct sets of basis functions. The orthogonality of basis functions is defined by analogy with the orthogonality of basis vectors.

The Fourier version of our model is obtained from (4) by defining $\psi(i, j)$ as $\cos(i\ \phi_j\ \pi/l)$ (see Figure 2), where $\{\phi_j\}_{j=1,k}$ is a subset of k different integers from $\{1, 2, \ldots, l - 1\}$.

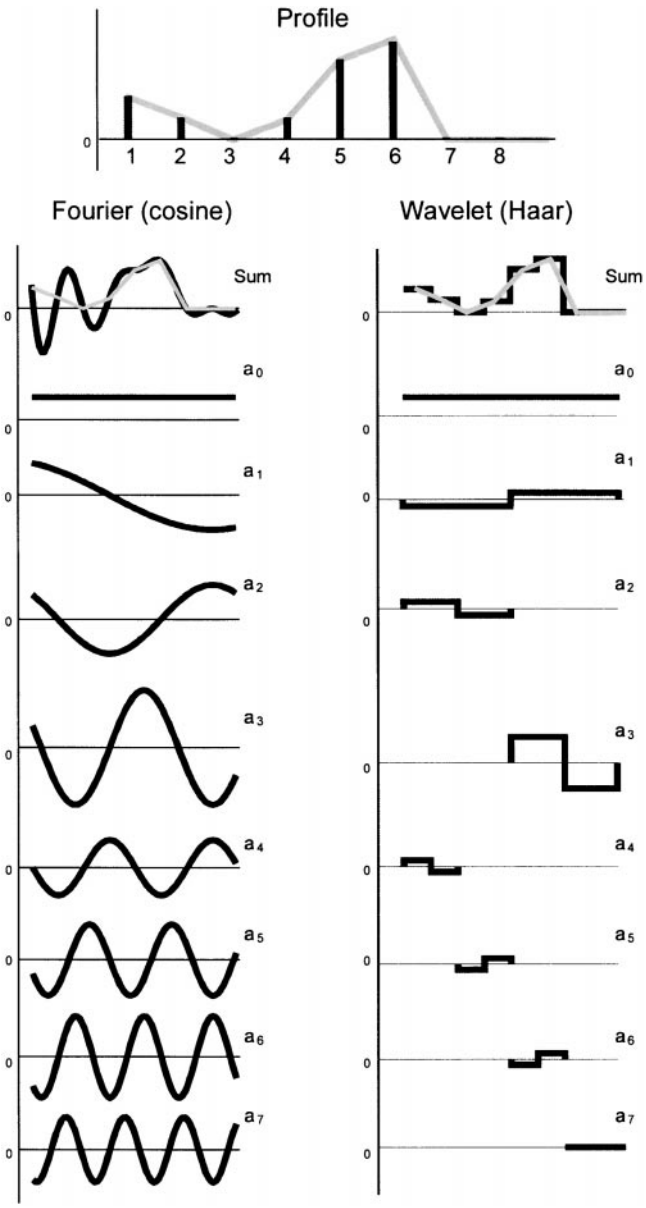

Figure 2.—An example of decompositions of the same discrete function ("profile") with discrete Fourier (cosine) and discrete wavelet (using the Haar mother wavelet) transforms. Although both methods give an exact fit to the target function with a complete number of coefficients, the resulting discrete Fourier decomposition is plotted as a continuous function mainly for aesthetic reasons. Note that the relative rate functions that we are using in this article are different from the common discrete Fourier and wavelet transforms in that the average value of our target function has to be equal to 1; due to this restriction, parameter a_0 is substituted with a constant set to 1 in both transforms.

$$f(i; a_1,a_2, \ldots, a_k)$$

$$= \left(1 + \sum_{j=1}^{k} a_j \cos\left\{\frac{i\phi_j\pi}{l}\right\}\right) \Big/ \left(1 + \sum_{p=1}^{k} \sum_{q=1}^{l} a_p \cos\left\{\frac{q\phi_p\pi}{1}\right\} l\right). \quad (5)$$

The choice of cosines as basis functions is motivated by

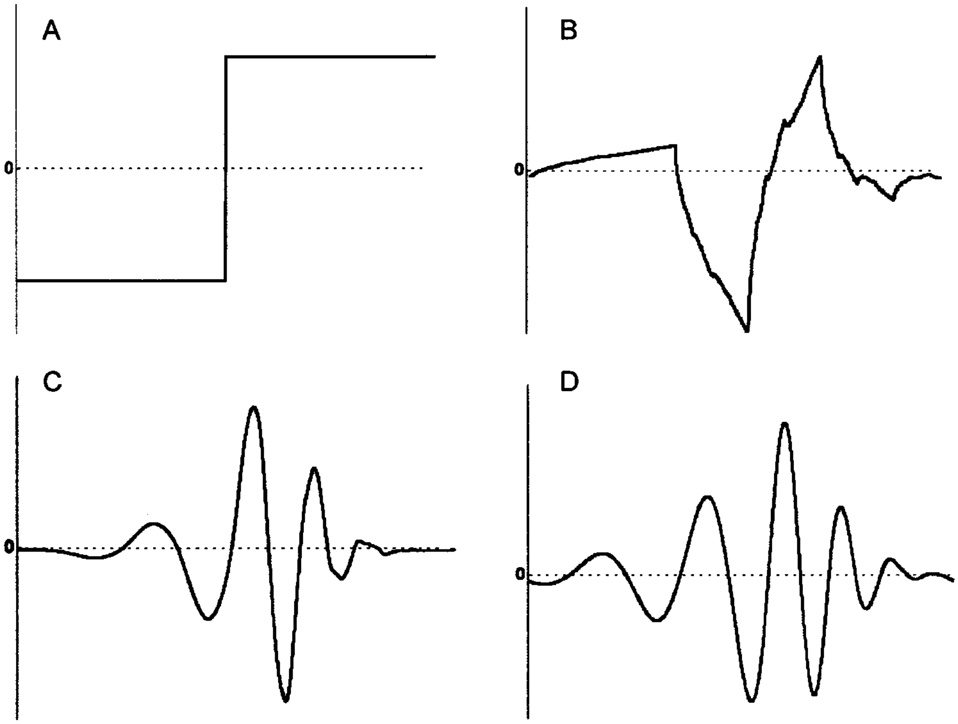

Figure 3.—The four types of the mother wavelets used in this article: (A) Haar wavelet (B) Daubechies 4 wavelet (C) Daubechies 12 wavelet, and (D) Daubechies 20 wavelet.

the well-known fact that cosines with period ranges from 1 to l form a valid basis of functions for exactly representing any discrete function (not necessarily periodic) on the interval between 1 and l (*e.g.*, see Bronstein and Semendiaev 1986). Further, by dropping the terms with the smallest absolute values of a_i's one can obtain an appropriate approximation of the original function.

The wavelet version of our model is obtained by defining $\psi(i, j)$ as one of the wavelet functions commonly used in discrete wavelet transforms (see Figure 2). Unlike basis functions in the Fourier series, wavelets are local functions, which is a convenient property for summarizing rate variation at different scales. The choice of wavelets as basis functions also allows for a considerable reduction in computation time required for maximizing the likelihood function compared to the corresponding computations under the Fourier model. In this work we use $\psi(i, j)$ based on Haar, Daubechies 4, Daubechies 12, and Daubechies 20 wavelets (see Figure 3; Daubechies 1988; Press *et al.* 1992).

A wavelet is a univariate real-valued function selected such that it vanishes outside a limited interval and has equal-sized areas below and above zero. The ultimate goal of a discrete wavelet transform is to represent and/or approximate a function given as a set of l values, where l is usually a power of 2. In our case l is the number of sites and the function that we are trying to approximate is the relative substitution rate at each site.

The wavelets are used for constructing an orthogonal basis of functions as described below for Haar wavelets. The basis function in the case of the simplest Haar transform has the shape of a step function (see Figures 2 and 3A) such that for a wavelet defined on interval [0, 1], the value is -1 on interval [0, 1/2], and $+1$ on interval [1/2, 1]. The basis functions for the Haar wavelet decomposition are generated by scaling and shifting the same step wavelet function such that exactly one function covers the complete interval [1, l], two basis functions cover exactly half of the complete interval (intervals [1, $l/2$], and [$l/2 + 1$, l], respectively), four basis functions cover one-fourth of the complete interval (intervals [1, $l/4$], [$l/4 + 1$, $l/2$], [$l/2 + 1$, $3l/4$], and [$3l/4 + 1$, l], respectively), and so on. The early wavelets can contribute to a large range (*e.g.*, left half-right half of the sequence) of variations, while successively more focused wavelets pick out smaller regions and eventually the individual sites. The basis functions with the smallest domain are most abundant (there are $l/2$ of them) and cover just two data points; such functions together cover the complete interval [1, l]. In addition to step functions, the complete set of Haar basis functions contains one scale function that has value 1 on interval [1, l]. The details of fast computation of a discrete wavelet decomposition can be found in Press *et al.* (1992). In real applications l is rarely an exact multiple of 2^n and thus the relative rates array must be padded with zeroes up to the nearest multiple of 2^n.

One may define the optimality of a set of basis functions $\psi(i, j)$ as the minimal number of functions from this set required to fit the data with a predefined minimum of the maximum-likelihood value. We suspect that there is no best set of basis functions that would enable one to obtain the most parsimonious description for any arbitrary data set. It is more likely that the Fourier version of the model would be more suitable for data sets with a pronounced periodicity of rates, while the wavelet version will work better for aperiodic and/or sparse rate variation profiles. In this article we illustrate the application of both.

Choosing the optimum rate variation model: When choosing among several alternative models for describing the same data, one intuitively searches for optimum trade-off between the goodness-of-fit of a model and economy of representation—the number of free parameters defined by the model. The general approach to such a problem is to *search* through the space of all possible models and make certain *comparisons* between competing models. We discuss these two steps, searching and comparison of models, separately.

Searching through the space of all possible models: Both the Fourier and wavelet models allow for the generation of a large number of rate variation profiles, ranging from a zero rate parameter profile to the most parameter rich ($l - 1$ relative rate variation parameters for a set of sequences of length l). Therefore, for each data set it would ideally be of interest to (*i*) order all possible rate variation profiles by their maximum-likelihood values, and (*ii*) use an objective scheme to decide what subset of rate variation parameters is needed for an optimum description of the data set.

The *searching* through the space of different profile models can be done in the following ways.

1. Backward elimination search. First, beginning with the "general" wavelet or Fourier profile using all $l - 1$ rate parameters, successive parameters may be removed one by one by iteratively deleting the parameter with the smallest absolute value and repeating the maximum-likelihood optimization under each reduced model. The process is repeated until the change in likelihood is "small," for example, when compared to a 5% chi-square critical value corresponding to the doubled difference between likelihoods of the "more complex" and "less complex" models (the likelihood-ratio test).
2. Forward selection search. One starts with the equal-rate model and increases the number of parameters by one, trying all possible combinations, until the increase in the likelihood is large, say, according to the likelihood-ratio test.
3. Markov chain Monte Carlo search. It is possible to substitute/supplement the maximum-likelihood analysis with a Bayesian analysis with a flat initial distribution using the Markov chain Monte Carlo (MCMC) technique for simultaneously searching for the best phylogenetic tree and the best rate variation model [*e.g.*, see "reversible jump" method by Green (1995) and Mau *et al.* (1996); the MCMC method is described in more detail in the next section]. We did not implement this strategy in the current study.
4. Heuristic search. There are numerous heuristics that can be used for the purpose. To decide the order in which cosine functions or wavelets should be added in stepwise computation of Fourier or wavelet profiles, we started by estimating all $l - 1$ parameters by the least-squares fitting of the cosine Fourier function to the observed profile calculated under the uniform rate model. Given the set of normalized numbers of substitutions per site, x_i's, estimated under the equal-rate model, one can compute the set of wavelet or Fourier parameter values that minimizes

$$\sum_{i=1}^{l} (f(i; a_1, a_2, \ldots, a_k) - x_i)^2,$$

where the number of rate variation parameters, k, is set to $l - 1$ at the beginning. One then proceeds by iteratively eliminating the parameters with the smallest absolutes, which give the smallest contribution to the resulting rate profile.

Comparison of alternative models: This can also be addressed in several different ways. For example, rival models, nested or nonnested, may be compared by using one of the approaches built on the likelihood analysis but penalizing parameter-rich models, such as the Akaike information criterion (AIC; Akaike 1974), computed as

$$\mathrm{AIC}_i = 2 \log L_i - 2 N_i, \tag{6}$$

where N_i is the number of parameters used in the *i*th model and L_i is the maximum-likelihood value obtained under that model. The idea behind this formula is to penalize an increase in the number of parameters if the addition of each new parameter increases the likelihood value by less than one unit of log-likelihood. The better the fit of the model to the data, the larger the AIC value will be. The AIC tends to favor parameter-rich models as the data sample size increases. A second popular criterion, the Bayesian information criterion (BIC; Schwarz 1978), has the definition

$$\mathrm{BIC}_i = 2 \log L_i - N_i \log n, \tag{7}$$

where n is the sample size. BIC usually tends to choose less parameter-rich models than AIC because in real data analyses log n is usually >2. In our case BIC does not seem to be very useful because under our model different sites within the same sequence are not identically distributed and the sample size is not well defined. If we would believe that in this case the sample size is 1 (as suggested by Z. Yang, personal communication), BIC becomes equivalent to the maximum-likelihood value itself and automatically chooses the most parameter-rich model.

Next, one can use chi-square approximation for distribution of doubled difference between likelihoods of "complex" and "simple" models (the Cox test). This approach gives results very similar to those of AIC and can be misleading for small data sets because chi square distribution may not be an appropriate approximation for distribution of the test statistic. To correct this potential problem with distribution of the test statistic, one can use a parametric bootstrap to estimate distribution of the likelihood-ratio statistic in the Cox test (Goldman 1993); this distribution can then be used for deciding whether the likelihood under the more complex model is significantly greater than that under the simpler model. Finally, the optimum number of parameters may be decided with the MCMC approach within the Bayesian framework (Green 1995), although we did not implement this approach in our study.

In this study we illustrated application of the majority of the above approaches of model comparison while leaving out only Green's (1995) test.

Tests of biological hypotheses with the Fourier/wavelet model: For many biological investigations it is of interest to identify regions within a protein alignment that have evolved significantly slower or faster than the average of the protein. Our model allows for such a test by generating confidence intervals around the relative replacement rates in each site of a protein alignment. It also allows for testing the homogeneity of replacement rates among sites within a protein and testing the identity of replacement rate profiles between two or more sets of homologous proteins. The described methods represent a straightforward application of classical testing theory (*e.g.*, Kendall 1956).

First, we outline the test for homogeneity of replacement rates. Denote a set of the maximum-likelihood estimates of k rate variation parameters by a row vector, $\hat{\mathbf{a}}$. To test the null hypothesis that the replacement rate is the same across sites of a given protein (that is, $\mathbf{a} = \mathbf{0}$ where $\mathbf{0}$ is a zero vector), we need to compute the Hessian matrix $\mathbf{H}$ at the point $\boldsymbol{\Theta} = \{\hat{\mathbf{x}}, \hat{\mathbf{a}}\}$. (Vector $\hat{\mathbf{x}}$ contains the maximum-likelihood estimates of the tree branch lengths and the rest of the model parameters not included in $\hat{\mathbf{a}}$.) Element H_{ij} of the matrix $\mathbf{H}$ is defined as the second partial derivative of the logarithm of the likelihood function taken with respect to θ_i and θ_j evaluated for a specified set of parameter values. The negated Hessian matrix evaluated at the maximum of the likelihood surface is known to asymptotically tend toward the inverse of the variance-covariance matrix for the vector of the maximum-likelihood estimates of model parameters (*e.g.*, see Edwards 1972). Because the maximum-likelihood estimates are asymptotically normally distributed (*e.g.*, see Kendall 1956), under the null hypothesis of homogeneity of replacement rates across protein sites, the quadratic form

$$x_0 = \hat{\mathbf{a}}\, \mathbf{V}_a^{-1}\, \hat{\mathbf{a}}^{\mathrm{t}} \tag{8}$$

asymptotically follows a χ^2-distribution with k d.f., where $\mathbf{V}_a$ is a submatrix of the complete variance-covariance matrix corresponding to the vector of Fourier parameters and $\mathbf{V}_a^{-1}$ is its inverse; the superscript t indicates the transpose of a vector (this test is a special case of the Wald test). Unfortunately, because of the restrictions on the values of rate variation parameters under the wavelet and Fourier models (the sum of all relative rate values across sites should always be equal to l for a data set with l sites, and negative values of relative rates are not allowed), depending on the expected values of the relative rates in the data set under analysis, the actual distribution of the maximum-likelihood estimates may significantly deviate from normal and the utility of the described test is limited.

Alternatively, one can use the likelihood-ratio test for the same purpose. The random variable

$$\lambda = -2 \log \frac{\sup L(\mathbf{a} = \mathbf{0}; \mathbf{x})}{\sup L(\mathbf{a}; \mathbf{x})} \tag{9}$$

asymptotically follows a χ^2-distribution with k d.f. under the null model, provided that the more general model has exactly k extra rate variation parameters (a likelihood-ratio test). The numerator and denominator in (9) represent the maximum-likelihood values under the equal-rate and the wavelet/Fourier rate variation models, respectively. As with the Wald test, the actual distribution of the likelihood-ratio test statistic may not be close to the asymptotic χ^2-distribution for small data sets.

We next describe the comparison of replacement rate profiles from two sets of homologous protein sequences, $\mathbf{S}_1$ and $\mathbf{S}_2$. The independent wavelet or Fourier parameter estimates under the same rate variation model (*i.e.*, under models with the identical number and type of rate variation parameters), $\hat{\mathbf{a}}_1$ and $\hat{\mathbf{a}}_2$, respectively, for these two data sets are evaluated as

$$x_1 = (\hat{\mathbf{a}}_1 - \hat{\mathbf{a}}_2)(\mathbf{V}_{a1} + \mathbf{V}_{a2})^{-1}\, (\hat{\mathbf{a}}_1 - \hat{\mathbf{a}}_2)^{\mathrm{t}}, \tag{10}$$

which approximately follows a χ^2-distribution with k d.f. under the null hypothesis $E(\hat{\mathbf{a}}_1) = E(\hat{\mathbf{a}}_2)$, where $\mathbf{V}_{a1}$ and $\mathbf{V}_{a2}$ are the variance-covariance matrices for the vectors $\hat{\mathbf{a}}_1$ and $\hat{\mathbf{a}}_2$, computed from data sets $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively, in the same manner as described for $\mathbf{V}_a$ above. Once again, the actual distribution of rate parameter estimates appears to deviate significantly from normal for data sets with significantly nonuniform rate profiles, so one should exercise caution in applying this test to real data. This test statistic has a likelihood-ratio test statistic counterpart, which can be constructed by analogy with Equation 9.

Computing "confidence intervals" of relative substitution rates with the MCMC technique: In this study we applied MCMC technology to compute an analog of confidence intervals for relative rate profiles for a predefined tree topology. We used in this computation a fully parameterized rate variation model, but the procedure is obviously applicable to models with a more moderate number of rate variation parameters. The idea in this case is to use the property of MCMC analysis that the frequencies of model parameter values encountered in a well-designed random walk through parameter space, given a fixed tree topology, give an estimate of the posterior probabilities of these parameter values. In this way, observing the variation of the intermediate profiles in the MCMC random walk, one can compute intervals for rate parameters including the middle 95% of a posterior probability distribution with a flat prior, which we loosely refer to below as "95% confidence intervals."

In MCMC technique the likelihood function is computed exactly in the same way as it is done in the maximum-likelihood analysis, but instead of numerical maximization of the likelihood function, one simulates a random walk through the space of parameter values and, optionally, tree topologies (in our case we limit our analysis to a predefined tree topology, so the tree topology was not changed). The core iteration step of the MCMC calculation samples a new point in the parameter space, $\boldsymbol{\Theta}^*$, given the current value of $\boldsymbol{\Theta}$ and likelihood values at both points. Depending on the ratio of likelihood values at points $\boldsymbol{\Theta}^*$ and $\boldsymbol{\Theta}$, the system accepts the new point or rejects it, staying in the old one. The random walk can be designed in an infinite number of ways: the only restriction to the simulation design provided by MCMC theory (see Hastings 1970; Mau *et al.* 1996) is that the Markov chain describing the transition between states $\boldsymbol{\Theta}^*$ and $\boldsymbol{\Theta}$ be time reversible. Time-reversibility is defined as the requirement that the following equation hold true: $\Pi^* P(\boldsymbol{\Theta}^* \mid \boldsymbol{\Theta}) = \Pi P(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^*)$, where $P(\boldsymbol{\Theta}^* \mid \boldsymbol{\Theta})$ and $P(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^*)$ are the probabilities of transitions between states $\boldsymbol{\Theta}^*$ and $\boldsymbol{\Theta}$ in the Markov chain corresponding to the random walk, and Π^* and Π are the probabilities of states $\boldsymbol{\Theta}^*$ and $\boldsymbol{\Theta}$ in the posterior distribution that we are trying to estimate.

Following suggestions formulated by Hastings (1970), we used the following probability of *accepting* state $\boldsymbol{\Theta}^*$ being at state $\boldsymbol{\Theta}$:

$$P(\boldsymbol{\Theta}^* \mid \boldsymbol{\Theta}) = \alpha\ (\boldsymbol{\Theta}^* \mid \boldsymbol{\Theta})\ \min(1, L(\boldsymbol{\Theta}^*, T)/L(\boldsymbol{\Theta}, T)), \tag{11}$$

where $\alpha\ (\boldsymbol{\Theta}^* \mid \boldsymbol{\Theta})$ is the conditional probability of sampling $\boldsymbol{\Theta}^*$ given $\boldsymbol{\Theta}$. We designed the simulation such that the probabilities $\alpha(\boldsymbol{\Theta}^* \mid \Theta)$ and $\alpha(\boldsymbol{\Theta} \mid \Theta^*)$ were equal. The second multiplier in expression (11), $\min(1, L(\boldsymbol{\Theta}^*, T)/L(\boldsymbol{\Theta}, T))$, is the conditional probability of accepting point $\boldsymbol{\Theta}^*$ given that it is already sampled.

By definition, the value of the posterior distribution, Π, at point $\boldsymbol{\Theta}$ is a product of the likelihood $L(\boldsymbol{\Theta}, T)$, and the prior probability of observing the current $\boldsymbol{\Theta}$ and T, divided by the prior probability of observing the current data set. Assuming lack of a prior knowledge about the parameter/tree distribution (a uniform prior), we have $\Pi^*/\Pi = \mathrm{L}(\boldsymbol{\Theta}^*, T)/L(\boldsymbol{\Theta}, T)$, and therefore the time-reversibility condition, $\Pi^* P(\boldsymbol{\Theta}^* \mid \boldsymbol{\Theta}) = \Pi P(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^*)$, is indeed satisfied.

We organized our MCMC computation in the following way. The random walk always started at the point of the maximum of likelihood. Each iteration of the simulation included an update cycle through all parameters, updating one parameter at a time and each time deciding whether to accept or reject the updated value; the tree topology was not changed. For each parameter, θ_i, we defined the maximum absolute change of this parameter, δ_i, and allowed for its single stochastic update (usually δ_i was set to be equal to half of the absolute value of the maximum-likelihood estimate of θ_i). More precisely, first, an equiprobable random decision to decrease or increase the parameter value was made. Second, the absolute amount of change was determined by drawing a random number from a uniform distribution defined at the interval between 0 and δ_i. If the resulting value of δ_i was inadmissible (for example, the value of the relative rate became negative), δ_i retained the old value, and next the parameter was updated. [The latter step was essential to ensure the condition $\alpha(\boldsymbol{\Theta}^* \mid \boldsymbol{\Theta}) = \alpha(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^*)$—otherwise the probability of moving in the direction of the boundary of the region of allowed values is not generally equal to the probability of moving in the opposite direction.] For the sake of computational efficiency, under the unrestricted-rates model, the change in the value of one of the relative rate parameters was compensated by an equivalent change in the opposite direction in one of the other randomly chosen relative rate parameters. In this way we avoided the problem of renormalizing all relative rates after each change and had to recompute the likelihood values for only the two sites affected. If the new value of δ_i was admissible, the new point, $\boldsymbol{\Theta}^*$, was accepted with probability $\min(1, L(\boldsymbol{\Theta}^*, T)/L(\boldsymbol{\Theta}, T))$. Once all parameters were updated in this way, the resulting state of the system, $\boldsymbol{\Theta}$, was saved. The confidence intervals for relative substitution rates at each site were computed from 10,000 saved intermediate states of the random walk; the results of the first few iterations that preceded reaching the stationary state of the random walk were discarded.

Example 1. Variable regions of human immunoglobulins: To illustrate the application of the model, we analyzed two sets of human immunoglobulin light chain variable region protein sequences. The first set contained seven light κ chain variable region (V_κ) genes, representing three predominant V_κ subgroups (indicated in parentheses): O18 (I), L23 (I), L11 (I), A23 (II), O11 (II), L2 (III), and A27 (III) (Schäble and Zachau 1993). The second set contained seven light λ chain variable region (V_λ) genes, representing three frequently expressed V_λ families (indicated by the first digit): 1c, 1e, 2b2, 2d, 3a, 3e, and 3j (Williams *et al.* 1996). The nucleotide sequences of these genes were obtained from VBASE (Tomlinson *et al.* 1996) and translated to amino acid sequences using the MEGA computer program (Kumar *et al.* 1993).

For each data set the wavelet and Fourier parameters were estimated as follows: First, a neighbor-joining tree

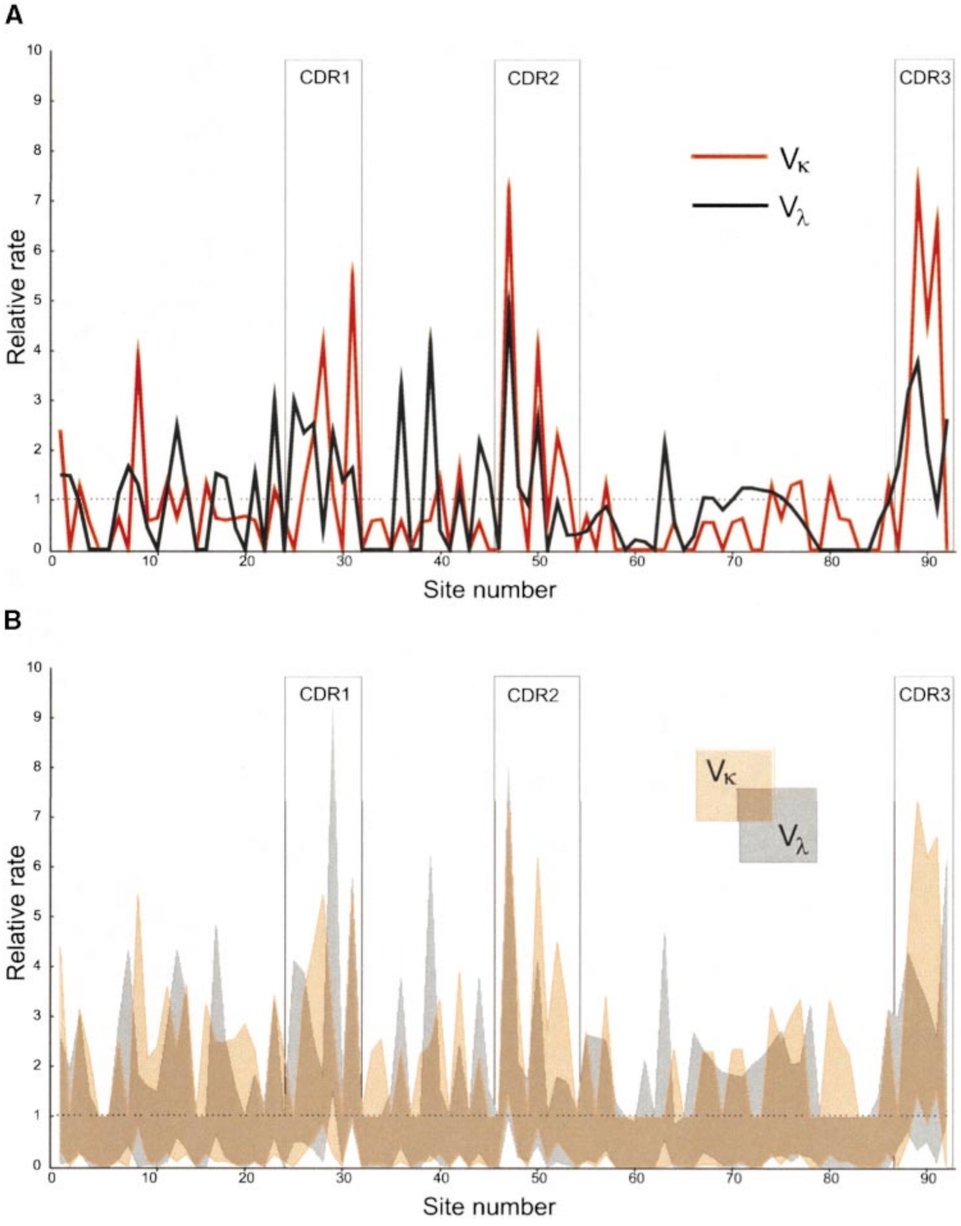

Figure 4.—(A) Relative substitution rates in variable domains of immunoglobulin light chains calculated with the maximum-likelihood analysis. (B) "Confidence intervals" based upon the middle 95% of a posterior probability distribution with a flat prior distribution (calculated by MCMC simulation) for the IgV_κ and IgV_λ data sets under the unrestricted-rates model. Also shown are the positions of CDRs, which are responsible for specific recognition of an antigen by antibody and are known to be especially variable.

(Saitou and Nei 1987) was computed from the aligned protein sequences using a Poisson model (Zuckerkandl and Pauling 1965) of amino acid replacement for estimating pairwise distances between sequences; the resulting unrooted phylogenetic tree was then used for maximizing the likelihood value under each model. (Note that in this application we computed substitution of the rate profiles for a neighbor-joining tree that is not necessarily the same as the maximum-likelihood tree; in work to be presented elsewhere we shall combine the search for the maximum-likelihood tree with estimation of the parameters under the Fourier / wavelet models.)

We used the heuristic method for ordering parameters in both models: starting with the equal-rate model, we computed the substitution rate profile and fit this profile with wavelet and Fourier functions, ordering the resulting parameter values for each function by decreasing absolute value. Once the parameters were ordered, we performed a series of maximum-likelihood analyses starting with the equal-rate model and serially adding wavelet or Fourier parameters as ordered in the previous step, beginning with the parameters of largest absolute value. Because addition of a new parameter to a Fourier function changes the values of the Fourier function in each site of the profile, while the analogous addition to a wavelet function changes values only in a subset of sites, the computation under the wavelet model is generally much faster. We encountered no difficulty in performing estimation while adding wavelet or Fourier parameters up until their maximum number: 91 for immunoglobulin sequence alignments of length 92. The resulting relative substitution rate profiles for the most parameter-rich variants were identical under the

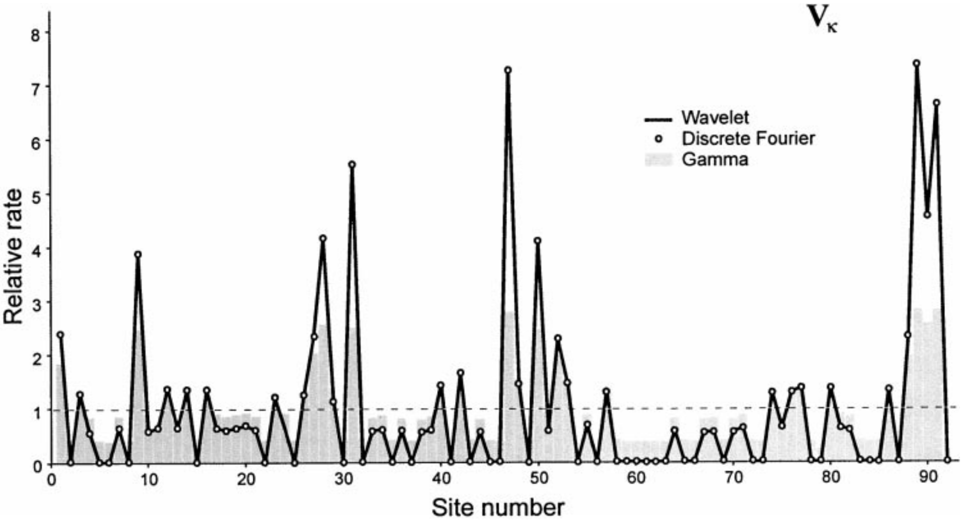

Figure 5.—Comparison of relative rate profiles for the IgV_κ data set obtained with the 91-parameter Fourier model, the 91-parameter Haar wavelet model, and the discrete gamma model with eight categories. The first two profiles are identical. The profiles recovered with the different models are generally very similar, although the discrete gamma model indicates less extreme rate heterogeneity.

Fourier model and wavelet models with the four different mother wavelet basis functions (see Figure 3) and corresponded to the profile obtained under the unrestricted-rates model. The 95% confidence intervals for the relative rates were computed under the unrestricted-rates model for both data sets.

The replacement-rates profiles are shown in Figure 4A and the corresponding confidence intervals are shown in Figure 4B. The regions of high replacement rate coincide with complementarity-determining regions (CDRs), which are the sites of antigen-antibody interaction. The regions of low replacement rate correspond to framework regions (FRs). For each data set the null hypothesis of rate constancy is rejected ($P < 0.05$) because the confidence intervals exclude rate 1 for at least one of the alignment sites. In contrast, we were unable to reject the null hypothesis that the patterns of rate variation are identical for these two IgV data sets, despite some differences between them. For example, the region between CDR2 and CDR3 (FR3) is slightly more variable in V_λ sequences than in the V_κ sequences.

The starting (equal-rate model) log-likelihood values were −885.06 and −896.67 for V_κ and V_λ data sets, respectively, while the final log-likelihood values of the most parameter-rich profiles were −794.14 and −804.33, respectively. Therefore, for both data sets the maximum-likelihood values under the unrestricted-rates model are >90 units of log-likelihood larger than under the equal-rate model. Using the PAML package (Yang 1998), we analyzed the same data sets under the discrete gamma model with eight rate categories. The estimated shape parameters of the gamma distribution were 0.99863 and 0.96650 for the V_κ and V_λ data sets, respectively, which corresponded to log-likelihood values of −873.73 and −885.13, respectively. Therefore, the maximum-likelihood values under the discrete gamma model are ~80 units of log-likelihood smaller than the corresponding values under the unrestricted-rates model. The profile obtained under the discrete gamma model with eight discrete categories for V_κ sequences is compared to the wavelet and Fourier model profiles for the same data set in Figure 5.

We performed a series of comparisons of different models of rate variation in terms of the optimum (maximum) AIC value. The data showed similar patterns among the different basis functions, so here we present only data for the Haar basis functions in the wavelet decomposition of the V_κ and V_λ data sets. For the V_κ data set the optimum value of AIC was −1705.51 and corresponded to 37-parameter Haar wavelet function (see Figure 6), while the AIC value for the V_κ data set under the discrete gamma model was −1771.47. For the V_λ data set the optimum value of AIC was −1724.85 and corresponded to 37-parameter Haar wavelet function, while the AIC value under the discrete gamma model was −1794.26. Therefore, in both of these analyses, AIC favored the wavelet/Fourier model over the discrete gamma model. We obtained essentially the same results as with AIC with the Cox test, assuming that the asymptotic distribution of the test statistic is valid (data not shown). This similarity of results between AIC and the Cox test is not completely unexpected, because both tests use virtually the same test statistic and the same assumption of the asymptotic distribution of the test statistic.

To verify the assumptions concerning the distribution of the Cox statistic (logarithm of the maximum-likelihood value under the discrete gamma model subtracted from the logarithm of the maximum-likelihood value under the wavelet model), we implemented the parametric bootstrap version of the Cox test as described by Goldman (1993). In this test we simulated data using the parameter estimates obtained under the discrete gamma model with eight discrete categories of rates; the generated sequences were assumed to exactly follow the discrete gamma model. For each of 100 data sets generated in this fashion we computed the maximum-

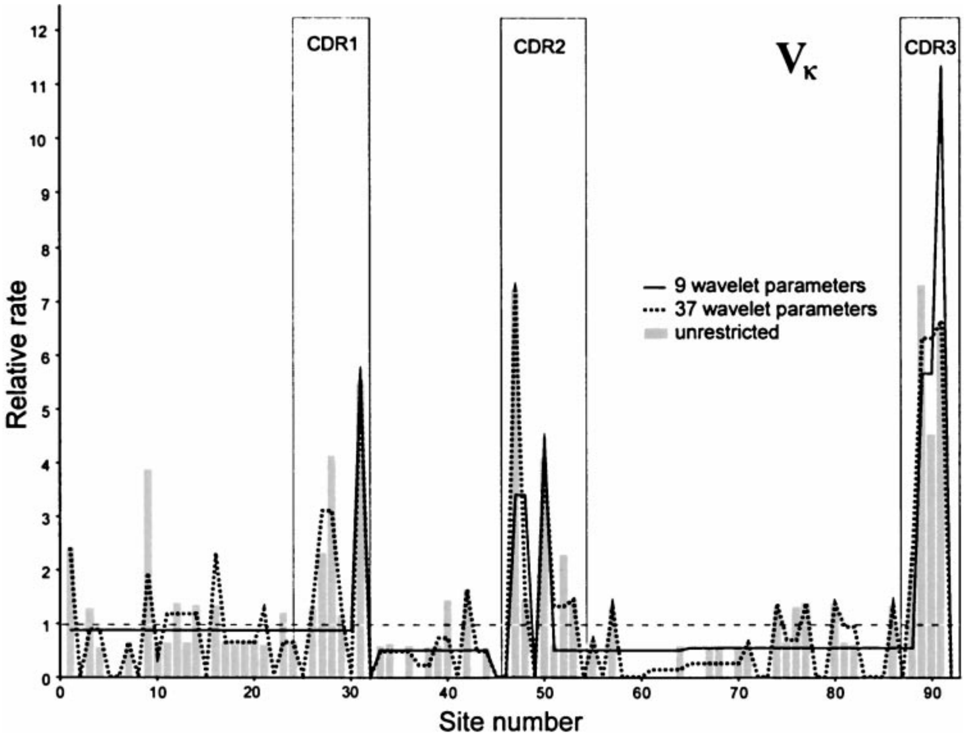

Figure 6.—Comparison of relative substitution rate profiles computed under 9-parameter, 37-parameter, and 91-parameter Haar wavelet models for the IgV_κ data set. The profile computed under the 37-parameter model appears to have most of the essential features of the profile reflected by the "full" 91-parameter model; the 9-parameter profile correctly reflects positions of three hypervariable regions CDRs but lacks details in the relatively slowly evolving "framework regions."

likelihood value under the wavelet model with 80 parameters and under the "true" discrete gamma model. The estimated distribution of the test statistic is shown in Figure 7; clearly, the probability of generating the observed or higher value of the test statistic observed for the data set V_κ is high under the null model and therefore the difference between the maximum-likelihood values for the wavelet/Fourier model and for the discrete gamma model is not significant for the V_κ data set. We conjecture that the situation is likely to change when a larger number of sequences is considered.

To see if the estimates of relative rates are sensitive to the assumed model of amino acid substitutions, we also compared estimates of relative rates obtained under

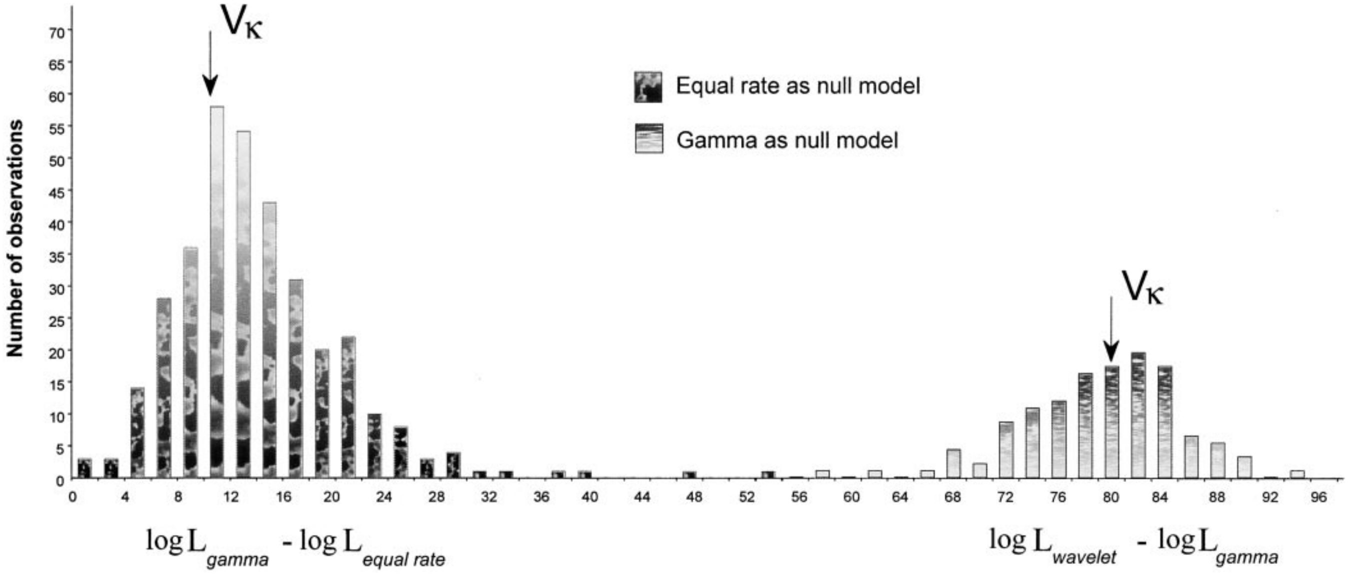

Figure 7.—Estimated distribution of the difference of log-likelihoods under the 80-parameter Haar wavelet model and the discrete gamma model with eight discrete categories of rates; the 100 data sets used for this computation were Monte Carlo generated using as the expected parameter values those estimated from the IgV_κ data set under the discrete gamma model. The arrow shows the value of the same statistic for the actual IgV_κ data set; according to this test, the value of the statistic for the IgV_κ data set is not significant.

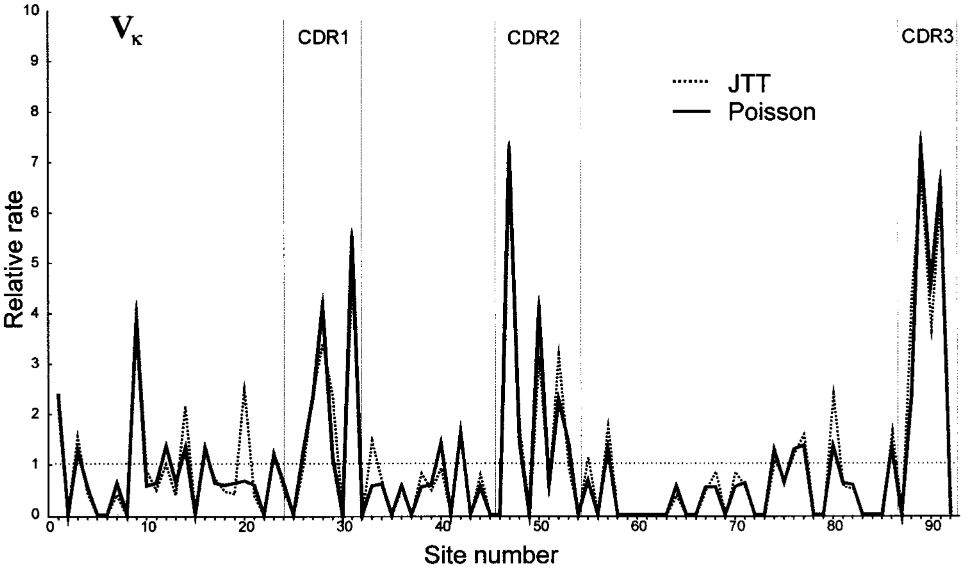

Figure 8.—Comparison of relative substitution rate profiles computed for the IgV_κ data set under the 91-parameter Haar wavelet model with Poisson (Zuckerkandl and Pauling 1965) and JTT (Jones *et al.* 1992) models of amino acid substitution. Despite the fact that the Poisson and JTT instantaneous substitution rate matrices are very different, the relative substitution rate profiles appear to be rather similar to each other.

the Poisson model with those computed under the "more realistic" JTT model for the V_κ data set (see Figure 8). The resulting profiles appeared to be very similar, at least for the data sets that we analyzed (see Figure 8).

Example 2. Drosophila alcohol dehydrogenase genes: Drosophila alcohol dehydrogenase (ADH) has been the focus of much interest among evolutionary biologists (Sullivan *et al.* 1990), and the observations and inferences made from one species group have often been compared to those made at others. Here we compare the patterns of rate variation observed at three monophyletic species clusters: the melanogaster group, the repleta group, and the Hawaiian group. Each clade is between 6 and 15 million years old; the repleta group and the Hawaiian group diverged from each other approximately 32 million years ago, and each diverged from the melanogaster group approximately 38 million years ago (Takezaki *et al.* 1995). The extent to which different evolutionary forces may be operative within the three clades may be addressed by an analysis of rate variation using the wavelet/Fourier model.

While ADH presumably performs the same biochemical functions in each of the species groups, the genetic and population genetic contexts are known to differ among them. For example, the Hawaiian species exist in limited geographic ranges and have effective population sizes (N_e) much smaller than the N_e's of the globally distributed species of the melanogaster group (DeSalle and Templeton 1986; Ayala *et al.* 1996). Differences in N_e can affect the patterns of replacement inasmuch as mutation in certain regions of the ADH sequence may be only slightly deleterious yet efficiently removed by selection in populations with large N_e, while becoming effectively neutral and thus allowed to accumulate in populations with relatively small N_e (Ohta 1993). Furthermore, as the Hawaiian species have adapted to narrow and specific environments, niche-specific adaptations may have resulted in altered profiles of replacements within the clade. Also, the repleta group is known to contain several ADH duplications, with different copies having evolved developmental stage-specific functions (Fischer and Maniatis 1985; Russo *et al.*1995). While ADH in species groups containing only a single copy are of necessity constrained to embody selective requirements of the entire life history, repleta group ADH-encoding genes may have accumulated developmental stage specific replacements in regions left unvaried in the other groups.

It is therefore of interest to compare the patterns of ADH replacement rate variation among the various Drosophilid species groups. Dorit and Ayala (1995) characterized rate variation by fitting cubic splines (analogous to multinomial regression) to the profiles of amino acid replacements across the length of the ADH sequences. Here we present the wavelet/Fourier approach to the same problem. In Figure 9 the results of fitting the unrestricted-rates model to the ADH sequences of the three species groups are presented. Each model is normalized to provide an average replacement rate of one, even though the number of replacements among the different groups is very different. While the profiles of the Fourier models differ markedly among the three groups, the relative girth of the confidence bands greatly diminishes any statistical significance that may be ascribed to the differences. In the melanogaster group (Figure 9A) the confidence interval contains the entire horizontal line passing through the relative rate of 1 representing no rate variation. Thus the null hypothesis of rate homogeneity across all sites of this sequence cannot be rejected. By contrast, in the repleta and Hawaiian groups the confidence intervals lie outside this line at some sites, indicating statistically significant dif-

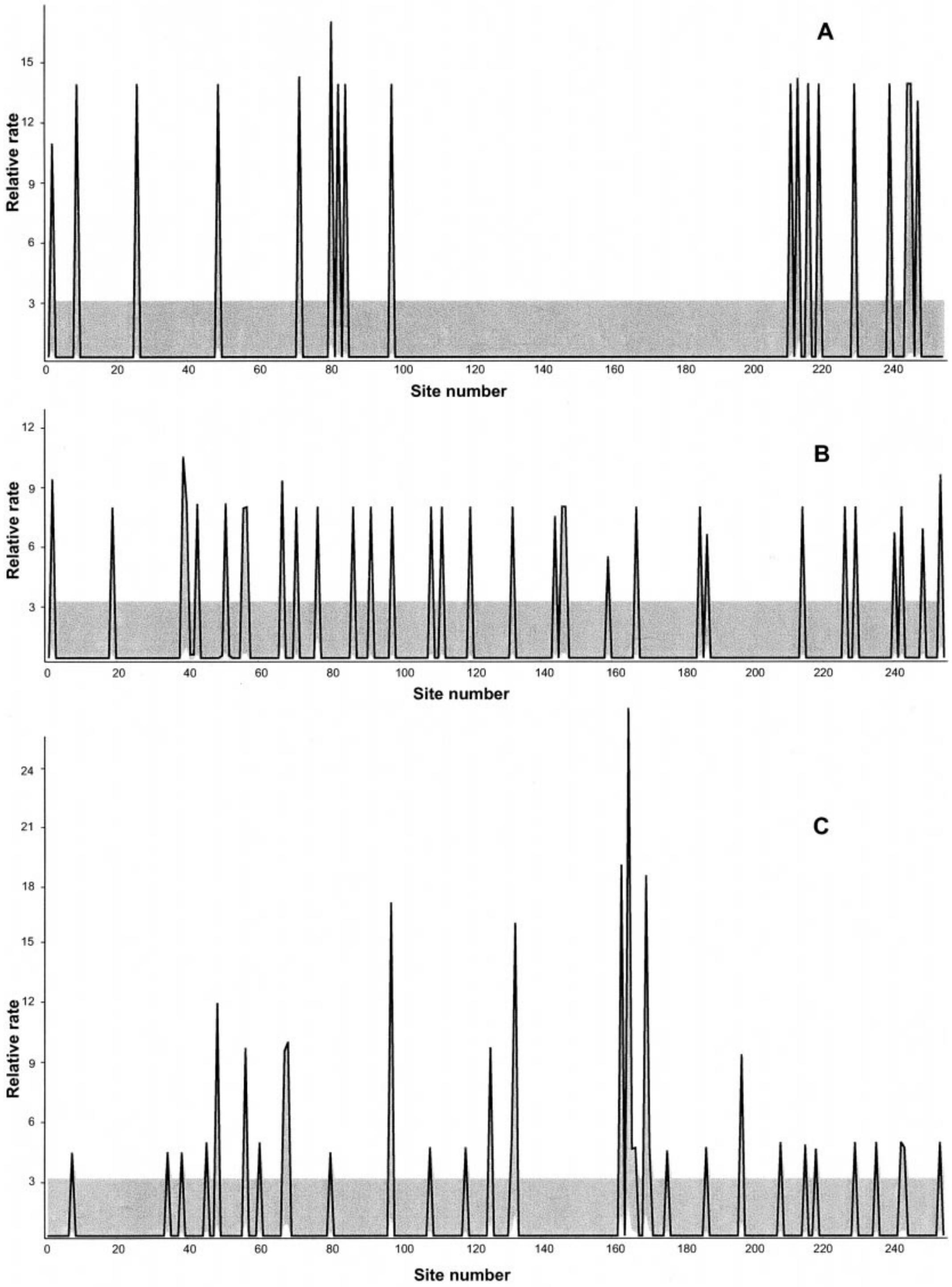

Figure 9.—Relative replacement rates and corresponding intervals based upon the middle 95% of a posterior probability distribution with a flat prior computed from alcohol dehydrogenase protein data sets representing three Drosophilid species groups under the unrestricted-rates model. The data sets contained ADH protein sequences from the melanogaster group [(A) *D. melanogaster*, *D. simulans*, *D. mauritiana*, *D. orena*, *D. erecta*, and *D. yakuba*], the repleta group [(B) *D. buzzatii*, *D. hydei*, *D. mayaguana*, *D. mettleri*, *D. mojavensis*, and *D. wheeleri*], and Hawaiian group [(C) *D. adiastola*, *D. affinidisjuncta*, *D. differens*, *D. mimica*, *D. nigra*, *D. picticornis*, and *D. silvestris*]. The data sets used in this computation are the same as in Dorit and Ayala (1995).

ferences in replacement rate at those sites. However, comparison of the three confidence intervals in Figure 9 reveals no significant difference between rate profiles among the three sequence sets. Thus the null hypothesis of equal patterns of replacements among all three species cannot be rejected by the Fourier/wavelet model analysis. Thus, despite the differing genetic and population genetic contexts in which their respective ADHs operate, no statistical difference in the relative rates across the length of the sequences within the melanogaster, repleta, or Hawaiian species groups has been detected.

DISCUSSION

Statistical consistency of parameter estimates under the unconstrained-rates and wavelet/Fourier models. The "big bang" model: Imagine a hypothetical data set with an infinite number of homologous DNA sequences diverged from a common ancestor according to a starlike tree under the Jukes-Cantor model (Jukes and Cantor 1969). The expected distance (number of substitutions per site) from the common ancestor is exactly the same for sites with the same number in different sequences (that is, sites with the same mean substitution rate). Next, consider a set of m genes of length l sampled from the infinite pool of genes. Because under the big bang model we are able to increase the number of sequences to infinity, all $l - 1$ relative rate parameters can be consistently estimated.

We conjecture that under the unconstrained-rates model combined with an arbitrary bifurcating tree, the relative rate parameter estimation has essentially the same properties as under the big bang model, except that the addition of new sequences leads to an increase in the number of the branch length parameters and is followed by a change in tree shape. Namely, an increase in the number of sequences to infinity should be associated with a reduction of the variance of relative rate parameters to an arbitrarily small value, while an increase in the total number of sites in the alignment should be followed by a reduction of the variance of the branch length estimates. According to this conjecture, one can decrease arbitrarily the variances of estimates of all model parameters by simultaneously increasing the number of sites and the number of sequences in the data set.

Strange shape of the relative rate confidence intervals under the unrestricted-rates model and the geometry of the space of the admissible parameter values: The confidence intervals for relative substitution rates computed with the MCMC analysis are far from being symmetrical with respect to the maximum-likelihood value. For invariant sites, the asymmetry arises because negative values of relative rates are not allowed and there always remains the possibility that the mean substitution rate at the site is in reality positive and yet the site by chance appears constant. For sites with high relative substitution rates, the maximum-likelihood estimates are often at the upper boundary of the confidence region. This is a consequence of the unusual geometry of the space of admissible parameter values under the unrestricted-rates model. Indeed, under this model the sum of the relative rates is restricted to l for a set of sequences of length l, and negative values of the relative rate parameters are not allowed. The space of admissible values of sets of relative rate parameters has the shape of a multidimensional pyramid, with the base corresponding to low values of relative rate parameters and the tip corresponding to large values. Further, the proportion of admissible random sets of the relative rate parameter values in the complete manifold of admissible relative rate sets rapidly decreases as we increase the relative rate of one of the parameters in the set. As a result, when the maximum-likelihood value of *one* of the relative rate parameters is high, the set of relative rate parameters corresponds to a point close to the tip of the pyramid space of the admissible parameter values, so that most of the admissible values are situated below this point.

Detecting signal and noise in substitution rates: Lake (1998) introduced a method for describing substitution rate variation along genes and proteins that bears a resemblance to our Fourier model. Lake's method is based on the theory developed by Weiner (1948) for signal processing in the presence of noise. The original problem solved by Weiner followed from the necessity of transmitting a meaningful signal (human speech, for example) through communication channels that introduce stochastic noise. Weiner showed that given a prior knowledge of the statistical properties of signal and noise, it is possible to clean the transmitted signal from the noise. Lake suggested treating the substitution rates estimated along genes or proteins as an analog of the frequency of the transmitted signal plotted as a function of time in Weiner's original problem. Although stochastic processes certainly play an important role in the origin and fixation of mutations, we are hesitant to accept Lake's definitions of signal and noise for the substitution rate variation along genes and proteins based only on remote analogy. Consequently, we did not attempt filtering the relative substitution rate profiles in this study.

Computational resources: The wavelet/Fourier model is significantly more parameter-rich than earlier models, which leads to an increased computational cost. With most currently available computational resources it is probably not feasible to do a full-scale maximum-likelihood analysis under parameter-rich models for large sets of sequences. However, it is likely that as the computational resources improve, the more complicated models will become more practical and hence more popular.

The maximization of the likelihood function under the wavelet model took between 1.5 and 3 days for each of the two immunoglobulin data sets, and ~4 days for each of the alcohol dehydrogenase data sets on a Sun Enterprise 3000 with 4 × 250 MHz processors, running on the Solaris 6.2 operation system. Each computation was done using a single processor. MCMC computations were completed at the rate of 2500 iterations per day for each of the immunoglobulin data sets, and at an ~1.5 times slower rate for the ADH data sets (we used 10,000 iterations for each data set for computing the confidence intervals). Computations under the Fourier model were significantly slower for both analyses (usually 5–15 times as slow as the analogous computation under the wavelet model) and were therefore not performed for all data sets.

The wavelet model using different mother wavelets: The speed of numerical optimization under the wavelet model turned out to depend significantly on the mother wavelet: the difference in computation speed was fivefold at times, although the resulting profile under the most parameter-rich setup was always identical (data not shown). For example, the fastest optimization for the IgV_λ data set was attained using the Daubechies 20 mother wavelet (see Figure 3D), while the fastest optimization for IgV_κ was attained using the Haar wavelet (see Figures 2 and 3A). This indicates that for any given data set one may need to try several mother wavelets to find the fastest. The difference in computation speed is probably caused by differences in shape of the likelihood surface under the alternative wavelet models. Nevertheless, we note that the shape of the rate variation profile is unaffected by the choice of mother wavelet (data not shown).

Differences between the wavelet/Fourier model and earlier models: The major difference between the methods presented here and those available previously is that in our model we do not treat the rates as random. Instead, they are viewed as an unknown function of site and we try to estimate that function using a combination of basis functions. The wavelet/Fourier model characterizes a protein sequence as an ordered set of nonidentical sites that can have different mean replacement rates. The majority of the previously suggested models depict proteins as unordered sets of sites sampled from the same general distribution, although there are models, *e.g.*, see Felsenstein and Churchill (1996), where sites are not identically distributed. There are other important differences between the wavelet/Fourier model and the alternative models. One is that, while the gamma and hidden Markov chain models do allow for identification of a set of random variables, giving the best prediction of relative replacement rates at each site (*e.g.*, see Yang and Wang 1995), it is not immediately clear how to test the identity of replacement profiles for two data sets under these models. Such a test is immediately available with the wavelet/Fourier method. Another significant difference is that the Fourier/wavelet model allows for flexible choice of the number of parameters appropriate for analysis of a particular data set, while in the majority of the alternative models the number of free parameters is constant.

Availability of computer programs: The computer programs used in this study are available in C and the program for Fourier is also partially in MatLab (Math Works Inc.) from Tatyana Sitnikova and Pavel Morozov.

This article has benefited from numerous suggestions and comments provided by Drs. Andrew Clark, Yasuo Ina, Bret Larget, Masatoshi Nei, James J. Russo, Ziheng Yang, and Andrey Zharkikh, who read an earlier version of it. We are grateful to Dr. Ziheng Yang for providing the numerical optimization subroutines used in the C version of the project programs, and to Dr. Joseph Felsenstein for a fruitful discussion of the model. This study was partially supported by grants from the National Institutes of Health and the National Science Foundation to Masatoshi Nei and Robert K. Selander.

LITERATURE CITED

Akaike, H., 1974 A new look at the statistical model identification. IEEE Trans. Autom. Contr. **AC-19:** 761–763.

Ayala, F. J., C. D. Campbell and R. K. Selander, 1996 Molecular population genetics of the alcohol dehydrogenase locus in the Hawaiian drosophilid *D. mimica*. Mol. Biol. Evol. **13:** 1363–1367.

Bronstein, I. N., and K. A. Semendiaev, 1986 *Mathematics for Engineers*. Nauka, Moscow (in Russian).

Daubechies, I., 1988 *Wavelets*. S.I.A.M., Philadephia.

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*, edited by M. O. Dayhoff. National Biomedical Research Foundation, Washington, DC.

DeSalle, R., and A. R. Templeton, 1986 The molecular through ecological genetics of abnormal abdomen. III. Tissue-specific differential replication of ribosomal genes modulates the abnormal abdomen phenotype in *Drosophila mercatorum*. Genetics **112:** 877–886.

Dorit, R. L., and F. J. Ayala, 1995 ADH evolution and the phylogenetic footprint. J. Mol. Evol. **40:** 658–662.

Edwards, A. W. F., 1972 *Likelihood*. Cambridge University Press, Cambridge, UK.

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

Felsenstein, J., 1993 *PHYLIP: Phylogenetic Inference Package*. University of Washington, Seattle.

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.

Fischer, J. A., and T. Maniatis, 1985 Structure and transcription of the *Drosophila mulleri* alcohol dehydrogenase genes. Nucleic Acids Res. **13:** 6899–6917.

Fitch, W. M., and E. Margoliash, 1967 A method for estimating the number of invariant amino acid coding positions in a gene, using cytochrome c as a model case. Biochem. Genet. **1:** 65–71.

Fitch, W. M., and E. Markowitz, 1970 An improved method for determining codon variability in a gene and its application to the rate of fixations of mutations in evolution. Biochem. Genet. **4:** 579–593.

Golding, G. B., 1983 Estimates of DNA and protein sequence divergence: an examination of some assumptions. Mol. Biol. Evol. **1:** 125–142.

Goldman, N., 1993 Statistical tests of models of DNA substitution. J. Mol. Evol. **36:** 182–198.

Green, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82:** 711–732.

Hastings, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57:** 97–109.

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

Jones, D. T., W. R. Taylor and J. M. Thornton, 1992 The rapid generation of mutation data matrices from protein sequences. Comp. Appl. Biosci. **8:** 275–282.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules. pp. 21–132 in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Kelly, C., and G. A. Churchill, 1996 Biases in amino acid replacement matrices and alignment scores due to rate heterogeneity. J. Comput. Biol. **3:** 307–318.

Kelly, C., and J. Rice, 1996 Modeling nucleotide evolution: a heterogeneous rate analysis. Math. Biosci. **133:** 85–109.

Kendall, M. G., 1956 *The Advanced Theory of Statistics*, Ed. 3. Hafner, New York.

Kumar, S., K. Tamura and M. Nei, 1993 *MEGA: Molecular Evolutionary Genetics Analysis*. The Pennsylvania State University, University Park, PA.

Lake, J. A., 1998 Optimally recovering rate variation information from genomes and sequences: pattern filtering. Mol. Biol. Evol. **15:** 1224–1231.

Mau, W., M. A. Newton and B. Larget, 1996 Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Technical Report #961. Department of Statistics, University of Wisconsin, Madison, WI.

Ohta, T., 1993 Amino acid substitution at the ADH locus in Drosophila is facilitated by small population size. Proc. Natl. Acad. Sci. USA **90:** 4548–4551.

Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, 1992 *Numerical Recipes in C*. Cambridge University Press, New York.

Russo, C. A. M., N. Takezaki and M. Nei, 1995 Molecular phylogeny and divergence times of drosophilid species. Mol. Biol. Evol. **12:** 391–404.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **9:** 1119–1147.

Schäble, K. F., and H. G. Zachau, 1993 The variable genes of the human immunoglobulin κ locus. Biol. Chem. Hoppe-Seyler **374:** 1001–1022.

Schwarz, G., 1978 Estimating the dimension of a model. Ann. Stat. **6:** 461–464.

Strung, G., 1992 Wavelets. Am. Sci. **82:** 250–255.

Sullivan, D. T., P. W. Atkinson and W. T. Starmer, 1990 Molecular evolution of the alcohol dehydrogenase genes in the genus Drosophila, pp. 107–148 in *Evolutionary Biology*, edited by M. K. Hecht, B. Wallace and R. J. Macintyre. Plenum Press, New York.

Takahata, N., 1991 Overdispersed molecular clock at the major histocompatibility complex loci. Proc. R. Soc. Lond. B Biol. Sci. **243:** 13–18.

Takezaki, N., A. Rzhetsky and M. Nei, 1995 Phylogenetic test of the molecular clock and linearized trees. Mol. Biol. Evol. **12:** 823–833.

Tomlinson, I. M., S. C. Williams, O. Ignatovich, S. J. Corbett and G. Winter, 1996 *V BASE Sequence Directory*. MRC Centre for Protein Engineering, Cambridge, United Kingdom.

Wakeley, J., 1993 Substitution rate variation among sites in hypervariable region I of human mitochondrial DNA. J. Mol. Evol. **37:** 613–623.

Weiner, N., 1948 *Cybernetics*. The Technology Press, John Wiley & Sons, New York.

Williams, S. C., J.-P. Frippiat, I. M. Tomlinson, O. Ignatovich, M.-P. Lefranc *et al.*, 1996 Sequence and evolution of the human germline V_λ repertoire. J. Mol. Biol. **264:** 220–232.

Yang, Z., 1993 Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10:** 1396–1402.

Yang, Z., 1998 *Phylogenetic Analysis by Maximum Likelihood (PAML)*. Version 1.4. University College, London.

Yang, Z., and T. Wang, 1995 Mixed model analysis of DNA sequence evolution. Biometrics **51:** 552–561.

Zharkikh, A., 1994 Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. **39:** 315–329.

Zuckerkandl, E., and L. Pauling, 1965 Evolutionary divergence and convergence in proteins, pp. 97–166 in *Evolving Genes and Proteins*, edited by V. Bryson and H. J. Vogel. Academic Press, New York.

Communicating editor: J. Hey