An Evolutionary Space-Time Model with Varying Among-Site Dependencies

Adi Stern and Tal Pupko

Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Israel

It is now widely accepted that sites in a protein do not undergo independent evolutionary processes. The underlying assumption is that proteins are composed of conserved and variable linear domains, and thus rates at neighboring sites are correlated. In this paper, we comprehensively examine the performance of an autocorrelation model of evolutionary rates in protein sequences. We further develop a model in which the level of correlation between rates at adjacent sites is not equal at all sites of the protein. High correlation is expected, for example, in linear functional domains. On the other hand, when we consider nonlinear functional regions (e.g., active sites), low correlation is expected because the interaction between distant sites imposes independence of rates in the linear sequence. Our model is based on a hidden Markov model, which accounts for autocorrelation at certain regions of the protein and rate independence at others. We study the differences between the novel model and models which assume either independence or a fixed level of dependence throughout the protein. Using a diverse set of protein data sets we show that the novel model better fits most data sets. We further analyze the potassium-channel protein family and illustrate the relationship between the dependence of rates at adjacent sites and the tertiary structure of the protein.

Introduction

Probabilistic evolutionary models describe how characters (nucleotides, amino acids, or codons) evolve along a phylogenetic tree. However, one must choose the assumptions of the evolutionary models carefully in order to distinguish biologically relevant signals from random evolutionary noise. The closer the assumptions of the model to the biological reality, the more accurate and powerful the model is (Lio and Goldman 1998).

The main goal when using an evolutionary model is to ensure that it is expressive enough to describe the biological reality, yet does not over fit the observations (Nei and Kumar 2000). Many of the simplifications once assumed are now being relaxed, giving way to more powerful models. A classical example of oversimplification is the assumption of equal evolutionary rates at all sites of a protein. In proteins, the rates of evolution vary due to different selective constraints that are acting on different sites. Pioneered by Yang (1993), the majority of models now being used take into account the heterogeneity of evolutionary rates among sequence sites (Swofford et al. 1996; Yang 1996; Felsenstein 2001). Accordingly, the rate at each site is modeled as a random variable drawn from a specified prior distribution. By far, the most commonly chosen distribution for modeling rate variation across sites is the gamma distribution (Yang 1993, 1994).

Following the incorporation of rate heterogeneity into evolutionary models, more sophisticated models were developed to more accurately describe the distribution of rates among sites. For example, Gu, Fu, and Li (1995) suggested the gamma + invariant model whereby a proportion of the sites are invariant. This assumption was recently generalized in a model which assumes that the rates take upon a mixture of gamma distributions (Mayrose, Friedman, and Pupko 2005). Each of these models assimilates additional parameters representing different biological assumptions. Nevertheless, all of these models share one recurrent shortcoming: they all assume that each site evolves independently of the other sites. This oversimplification of the evolutionary model is perhaps as troubling as the assumption that all sites of a protein share the same evolutionary rate. A protein is composed of conserved regions as well as variable regions, pointing at the fact that the sites do not evolve independently. This assumption has been applied in methods which detect selective constraints acting on regions of proteins by using a sliding window approach (e.g., Fares et al. 2002). An alternative model-based approach was introduced by Yang (1995) and Felsenstein and Churchill (1996), who developed a model of evolution which takes into account a correlation between the evolutionary rates at adjacent nucleotides by using a hidden Markov model (HMM). These models can better account for linear regions of low rate and linear regions of high rate, constituting an important advance of evolutionary models. They have been shown to provide a better fit to DNA data and may improve site-specific rate inference (Felsenstein and Churchill 1996).

Yet when observing empirical rate distributions of known proteins, it seems as though the assumption whereby the protein is composed of conserved linear regions and variable linear regions does not necessarily hold true. In fact, although most sites are clustered together according to their rates, there are clusters of sites where the evolutionary rates oscillate between very high and very low values (fig. 1). Thus, the assumption whereby there is equal correlation between all the pairs of adjacent sites in a protein is invalid. The protein's 3-dimensional (3D) structure and function result from complex interactions between amino acids. For instance, the catalytic site of the protein is often composed of sites which are distant in the linear sequence of the protein. Thus, the level of correlation between the evolutionary rates of these linearly distant sites may be stronger than the correlation between the linear adjacent sites. Alpha helices and beta sheets may also display varying levels of correlation between rates. For instance, if one side of an alpha helix is more functionally important than the other side, then the pattern of rates will be cyclic. Furthermore, a pattern of buried versus exposed sites may also lead to varying levels of correlation. Thus, the 3D structure may impose independence of adjacent rates in the linear sequence.

Key words: evolutionary models, hidden Markov models, site-specific evolutionary rates.

E-mail: talp@post.tau.ac.il.

Mol. Biol. Evol. 23(2):392–400. 2006
doi:10.1093/molbev/msj044
Advance Access publication November 2, 2005

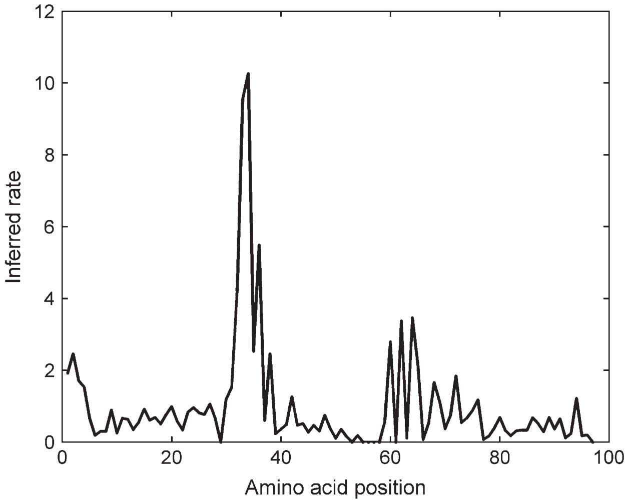


FIG. 1.—An example of a typical rate distribution (the K-channel protein family), as inferred by maximum likelihood (ML) using the Rate4Site program (Pupko et al. 2002). In most regions of the protein the rates are highly correlated, while between positions 58 and 65 the rates oscillate between low and high values.

In a step toward a more realistic evolutionary model, we propose a model that allows adjacent rates to be correlated at certain regions of the protein while allowing rate independence at other regions. This model implicitly takes into account the 3D structure of the protein by allowing flexibility of the correlation between the evolutionary rates. We hereby refer to this novel model as the *D* + *I* model, indicating dependence and independence, as opposed to the model described by Yang (1995), which we denote as the *D* model. We refer to a model where the sites are independent with a gamma distribution of rates across sites (Yang 1994) as the *I* model.

In order to evaluate which of the three models (*I*, *D*, and *D* + *I*) better fits protein data sets, we first applied all three models to a wide range of data sets. We showed that the *D* model is superior to the *I* model, as previously reported for DNA sequences, and the *D* + *I* model is superior to both. We then compared the performance of the three models in simulations, where the same pattern of performance emerged. Finally, we used the *D* + *I* model to analyze the potassium (K)-channel protein family and to infer site-specific evolutionary rates, revealing a particular pattern of correlation throughout the protein.

Theory

The Gamma Distribution

The most commonly used distribution for modeling rate variation at a single site is the gamma distribution (Swofford et al. 1996; Yang 1996). Its shape parameter, α, allows for different distribution shapes, making it a highly convenient distribution. In order to employ the gamma distribution, a discrete approximation is used (Yang 1994). The actual distribution is divided into *K* rate categories, such that all categories have equal prior probabilities (1/*K*). The mean of each category, denoted as $\bar{r}_i$, is used to represent all the rates within category *i*. The boundaries of each category are denoted as r^*_{i-1} and r^*_i (note that $r^*_0 = 0$ and $r^*_K = \infty$).

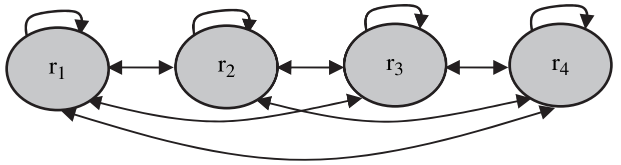


FIG. 2.—The *D* model. A representation of a Markov chain of evolutionary rates, where the gamma distribution is approximated by four categories. Arrows represent transitions between the rates at adjacent sites.

Modeling Rate Dependence Using an HMM

We model rate dependence by assuming that the evolutionary rates of the protein sites follow a stationary Markov chain (see figs. 2 and 3*A*), and each rate emits a corresponding column of the multiple sequence alignment (MSA). Note that given the rate assignment for site $n-1$, the rate distribution at site *n* is fully specified. An alternative model where the rate at site *n* depends on site $n+1$ instead of $n-1$ turns out to give identical results (Yang 1995).

An HMM is characterized by the transition probabilities between the states of the stationary Markov chain, by the initial probabilities of the hidden states, and by the emission probabilities, which represent the probability of the observations given assignments to the hidden states. Thus, in order to construct an HMM we must define these three sets of probabilities.

D Model: the Transition Probabilities

In the *D* model described by Yang (1995), the transition between rates at adjacent sites was modeled by assuming that the rates at adjacent sites follow a correlated bivariate gamma distribution. Since the theory behind this modeling is not trivial and since the theoretical background is given only briefly in the manuscript of Yang (1995), we will outline the essentials of the model.

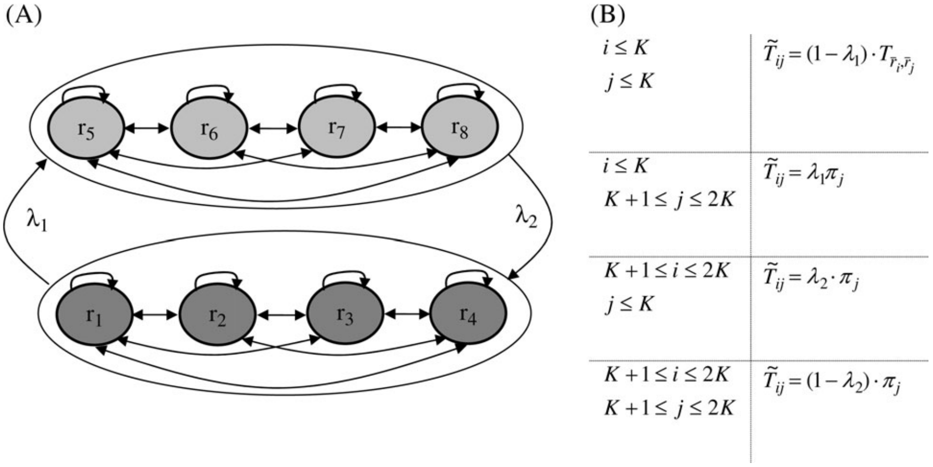
Let *T* represents the transition matrix between any two rate categories represented by $\bar{r}_j$ and $\bar{r}_k$:

$$T_{jk} = T_{\bar{r}_j,\bar{r}_k} = P(r^*_{k-1} < R_i < r^*_k | r^*_{j-1} < R_{i-1} < r^*_j) = \frac{P(r^*_{k-1} < R_i < r^*_k, r^*_{j-1} < R_{i-1} < r^*_j)}{p(r^*_{j-1} < R_{i-1} < r^*_j)}, \quad (1)$$

where R_{i-1} and R_i are random variables representing the hidden rates at any two adjacent sites $i-1$ and *i* along the protein and $1 \le j,k \le K$. (r^*_{j-1}, r^*_j) and (r^*_{k-1}, r^*_k) represent the boundaries of categories *j* and *k*, respectively, under a discrete approximation of the marginal distributions of the bivariate gamma distribution, both of which are univariate gamma distributions. The numerator of the fraction is computed by calculating the volume of a bivariate gamma distribution over the rectangle $(r^*_{k-1}, r^*_k) \times (r^*_{j-1}, r^*_j)$ (Yang 1995).

D Model: the Initial Probabilities

Because the gamma distribution is approximated by *K* equally probable categories, the initial distribution of the



$i \le K$ $j \le K$	$\tilde{T}_{ij} = (1-\lambda_1)\cdot T_{\bar{r}_i,\bar{r}_j}$
$i \le K$ $K+1 \le j \le 2K$	$\tilde{T}_{ij} = \lambda_1 \pi_j$
$K+1 \le i \le 2K$ $j \le K$	$\tilde{T}_{ij} = \lambda_2 \cdot \pi_j$
$K+1 \le i \le 2K$ $K+1 \le j \le 2K$	$\tilde{T}_{ij} = (1-\lambda_2)\cdot \pi_j$

FIG. 3.—The $D + I$ model. (*A*) A representation of a Markov chain of evolutionary rates, where the gamma distribution is approximated by four categories. Arrows represent transitions between the states. (*B*) The matrix $\tilde{T}$ represents the extended transition matrix between the rates given their dependence property, whereas T represents the transition matrix for the D model.

rates is simply a vector π of length K where $\pi(i) = 1/K\,(i = 1, \ldots, K)$. Note that the initial distribution π is a steady-state distribution, which fulfills the requirement:

$$\sum_{i=1}^{K} \pi_i T_{ij} = \pi_j. \tag{2}$$

This is easily verified because T is symmetric, and thus the sum of a column equals the sum of a row which equals one (as T is a Markov matrix).

The assumption that π is a steady-state distribution implies that the rate distribution does not change as we move along the Markov chain.

D Model: the Emission Probabilities

The emission probabilities of the D model are the likelihoods of the data at a position, given a certain rate assignment at this position. These probabilities are calculated using a postorder tree traversal algorithm (Felsenstein 1981). Note that we assume conditional independence of the data over sites given the rates, that is, once the sites are assigned their rates, each site evolves independently.

Parameters in the D model are $\theta = \{\alpha, \rho\}$. ρ is a measure of the correlation between adjacent sites and is a parameter of the bivariate gamma distribution. θ is maximized over the likelihood of the data (denoted as d):

$$\begin{aligned} L = P(d \mid \theta) &= p(d_1, \ldots, d_n|\theta) \\ &= \sum_{i_1=1}^{K} \cdots \sum_{i_n=1}^{K} \Bigg[P(R_1 = \bar{r}_{i_1}) \cdot P(d_1 \mid R_1 = \bar{r}_{i_1}) \\ &\quad \times \prod_{j=2}^{n} T_{i_{j-1},i_j}(\theta) \cdot P(d_j \mid R_j = \bar{r}_{i_j}) \Bigg]. \end{aligned} \tag{3}$$

For an elaboration of equation (3) see Appendix A.

The likelihood function can be calculated using the backward dynamic programming algorithm described in Durbin et al. (1998). Let $b_k(i) = P(d_{i+1}, \ldots, d_n|R_i = \bar{r}_k)$ be the probability of observing the partial data from sites $i + 1$ through n, given that site i is from category k. Then

$$b_k(i) = \sum_{j=1}^{K} T_{kj} b_j(i+1) P(d_{i+1} \mid R_{i+1} = \bar{r}_j) \tag{4}$$

with $b_k(n) = 1$.
The likelihood is thus

$$L = \sum_{k=1}^{K} \pi_k \cdot b_k(1) \cdot P(d_1 \mid R_1 = \bar{r}_k). \tag{5}$$

D + I Model: Incorporating Dependence and Independence

We propose a model that integrates between two types of situations: one where the rates are correlated, as described above, and the other where the rate at a site is independent of the previous rate. The model enables the rates to switch between these two types of situations by expanding the hidden states of the HMM. Thus, the number of possible hidden states is doubled. Whereas in the previous model, the hidden states could take upon any value between $\bar{r}_1$ and $\bar{r}_K$, now we rename and extend the hidden states to belong to two different sets: $S_d = (\bar{r}_1|\text{dependence}, \ldots, \ \bar{r}_K|\text{dependence})$ and $S_i = (\bar{r}_1|\text{independence}, \ldots, \ \bar{r}_K|\text{independence})$. Thus, each position has two hidden properties—its evolutionary rate and whether the rate is dependent or independent of the rate at the previous site. For simplicity, we denote $(\bar{r}_1|\text{dependence}, \ldots, \bar{r}_K|\text{dependence})$ as $(\bar{r}_1, \ldots, \bar{r}_K)$ and $(\bar{r}_1|\text{independence}, \ldots, \bar{r}_K|\text{independence})$ as $(\bar{r}_{K+1}, \ldots, \bar{r}_{2K})$. Figure 3*A* depicts this extended HMM.

D + I Model: Transition and Initial Probabilities

Let $\tilde{T}$ represents the transition matrix for the extended HMM, where the set of states is $S_d \cup S_i$. Two new parameters λ_1 and λ_2 represent the probabilities of a transition between S_d and S_i and vice versa, respectively (fig. 3*B*). We are left with the task of computing the initial state probabilities, denoted as ψ. We compute ψ via the assumption that ψ is a steady-state distribution (as in eq. 2 above):

$$\sum_{i=1}^{2K} \psi_i \tilde{T}_{ij} = \psi_j. \tag{6}$$

Together with

$$\sum_{i=1}^{2K} \psi_i = 1 \tag{7}$$

we have a set of linear equations which can be solved to obtain the values of ψ (note that the last equation in the set of equations derived from equation (6) is redundant due to the Markov matrix properties, leaving us with a total of $2K$ equations and $2K$ variables).

Unlike π, the initial probabilities $\psi_1, \ldots, \psi_K$ and $\psi_{K+1}, \ldots, \psi_{2K}$ are not equal.

$D + I$ Model: the Emission Probabilities

The likelihood function is now extended to incorporate the new states:

$$L = P(d \mid \theta) = \sum_{i_1=1}^{2K} \cdots \sum_{i_n=1}^{2K} \Big[P(R_1 = \bar{r}_{i_1}) \times P(d_1 \mid R_1 = \bar{r}_{i_1}) \times \prod_{j=2}^{n} \tilde{T}_{\bar{r}_{i_{j-1}}, \bar{r}_{i_j}}(\theta) \cdot P(d_j \mid R_j = \bar{r}_{i_j})\Big], \tag{8}$$

where the $2K$ states are the states in $S_d \cup S_i$.

Equations (4) and (5) can be changed accordingly to include the new states, and the summations are extended to include $2K$ different states.

Bayesian Estimation of Site-Specific Evolutionary Rates Using the D and $D + I$ Models

Estimating the rate at which a site evolves can be used as a means for inferring conserved and variable sites of a protein. Assuming independence between sites, it has recently been shown that an empirical Bayesian site-specific rate inference method is superior to a maximum likelihood (ML)-based approach (Mayrose et al. 2004). Here we study whether assuming the D or $D + I$ models can further improve inference.

In order to estimate the rate at each site, let us look at the conditional probabilities of the rates given the data. With the assumption of conditional independence it has been shown that the use of the conditional mean $\hat{r} = E(r|\text{data})$ as the predictor of the true rate (r) yields more precise inference than other predictors (Yang and Wang 1995):

$$\hat{r}_i = E(r_i \mid d) \cong \sum_{k=1}^{M} \bar{r}_k \cdot P(r_i \mid d) = \sum_{k=1}^{M} \bar{r}_k \cdot \frac{f_k(i) \cdot b_k(i)}{P(d)}, \tag{9}$$

where $M = K$ for the D model and $2K$ for the $D + I$ model. $P(d)$ and $b_k(i)$ are initially defined in equations (4) and (5), respectively, and are calculated as the ML estimators of the parameters of the D and $D + I$ models. $f_k(i) = P(d_1, \ldots, d_i, R_i = \bar{r}_k)$ is the joint probability of observing sites 1 through i where the rate at site i is from category k. $f_k(i)$ can be computed using a dynamic algorithm (Durbin et al. 1998):

$$f_k(i) = P(d_i \mid R_i = \bar{r}_k) \sum_{j=1}^{K} T_{jk} f_j(i-1) \tag{10}$$

with $f_k(1) = P(d_1, R_1 = \bar{r}_k) = \pi_k \cdot P(d_1|R_1 = \bar{r}_k)$.

Classification of Sites: Dependent or Independent

One novel use of the $D + I$ model may be in classifying the dependence property at each site to either S_d or S_i. It is appealing to focus on those regions where sites are classified as independent because these areas may point at a unique tertiary structure.

We use the maximum posterior estimator to classify site j to S_d or S_i:

$$P(R_j \in S_i \mid d) = \sum_{k=1}^{K} p(R_j = \bar{r}_k \mid d) = \sum_{k=1}^{K} \frac{f_k(i) \cdot b_k(i)}{p(d)}. \tag{11}$$

$P(R_j \in S_d \mid d)$ is calculated similarly, with the summation extending from $K + 1$ through $2K$. Site j will be classified as S_i if $P(R_j \in S_i \mid d) > P(R_j \in S_d \mid d)$ and to S_d otherwise.

Model Comparison

All analyses conducted in this study used the Jones-Taylor-Thornton (JTT) model of amino acid replacement (Jones, Taylor, and Thornton 1992). However, incorporating any of the three models into a desired nucleotide, codon, or amino acid substitution model is a trivial extension.

The likelihood ratio test (LRT) was used in order to test whether a specific model fitted a particular data set significantly better than another model. The LRT is applicable because all three models are nested: when $\lambda_1 = \lambda_2 = 0$, the $D + I$ model collapses into the D model, and when $\rho = 0$, the D model collapses into the I model. The $D + I$ model has two additional parameters (λ_1 and λ_2) as compared with the D model. Hence, the addition of these two parameters is statistically justified if the log-likelihood improvement is at least 2.995 ($P < 0.05$; chi square with 2 degrees of freedom). The D model has one additional parameter (ρ) as compared with the I model. Thus, the addition of this parameter is statistically justified if the log-likelihood improvement is at least 1.92 ($P < 0.05$; chi square with 1 degree of freedom). Although the models are nested, the use of the LRT may not be justified because of boundary problems (Anisimova, Bielawski, and Yang 2001). We thus also used the 2nd order Akaike Information Criterion (AIC) (Akaike 1974), defined as:

$$\text{AIC}_\text{C} = -2 \times \log L + 2p \cdot \frac{N}{N - p - 1}, \tag{12}$$

where p represents the number of free parameters and N represents the number of sequences in the data set. Using the AIC gave almost identical results as the use of the LRT.

We compared between the three models using 84 protein data sets (Aloy et al. 2001). For each data set, a phylogenetic tree was reconstructed using the neighbor-joining (NJ) algorithm (Saitou and Nei 1987) with pairwise distances calculated using the Jukes-Cantor distance. The parameters of

Table 1
AIC Scores and Maximum Log-Likelihood Values for the Analysis of Ten Data Sets Under the *I*, *D*, and *D* + *I* Models

Data Set[a]	SL[b]	NS[c]	AIC Score (log likelihood)			*P* Value (*D*, *I*)[d]	*P* Value (*D* + *I*, *D*)[d]
			I	*D*	*D* + *I*		
2ace	500	80	**79,033.16 (−39,515)**	**78,625.99 (−39,310)**	**78,535.29 (−39,263)**	**$<10^{-6}$**	**$<10^{-6}$**
1dxy	300	81	**70,862.50 (−35,430)**	**70,618.14 (−35,307)**	**70,517.18 (−35,254)**	**$<10^{-6}$**	**$<10^{-6}$**
1huh	200	83	**40,611.32 (−20,304)**	**40,484.14 (−20,240)**	**40,458.74 (−20,225)**	**$<10^{-6}$**	**0.0009**
1sgt	250	83	**35,441.14 (−17,719)**	**35,313.07 (−17,654)**	**35,279.47 (−17,635)**	**$<10^{-6}$**	**$<10^{-4}$**
6rsa	100	83	**12,709.67** (−6,353)	**12,709.02** (−6,352)	**12,708.88** (−6,350)	0.241	0.37
1rne	300	84	**44,380.77 (−22,189)**	**44,289.06 (−22,142)**	**44,253.16 (−22,122)**	**$<10^{-6}$**	**$<10^{-4}$**
1cle	500	85	**73,896.23 (−36,947)**	**73,451.84 (−36,723)**	**73,372.73 (−36,682)**	**$<10^{-6}$**	**$<10^{-6}$**
1mla	300	85	**74,750.24 (−37,374)**	**74,463.68 (−37,229)**	**74,439.63 (−37,215)**	**$<10^{-6}$**	**0.0009**
1quf	250	87	**33,056.61 (−16,527)**	**32,896.39 (−16,446)**	**32,887.89 (−16,439)**	**$<10^{-6}$**	**0.03**
1bro	250	88	**77,884.28 (−38,941)**	**77,545.14 (−38,770)**	**77,427.41 (−38,709)**	**$<10^{-6}$**	**$<10^{-6}$**

NOTE.—Values are shown in bold type if the AIC (LRT) score between the $D + I$ versus D or D versus I is lower (significant: $P < 0.05$).

[a] Data sets are referred to by their Protein Databank (Berman et al. 2000) ID.

[b] Sequence length.

[c] Number of sequences.

[d] P value between log-likelihood values of D ($D + I$) and I (D) models following LRT.

each model were optimized using the ML paradigm, and the gamma distribution was approximated using 16 categories.

Simulation Study

Simulations were used in order to examine the accuracy of site-specific rate estimates under different models. We simulated a given site with a specific “true” rate. An MSA is thus generated based on a vector of true rates. Subsequently, a rate for each column is inferred using the *I*, *D*, or *D* + *I* models. The closer the inferred rates to the true rates, the better the inference. For the simulations, one must determine the true rate in each site. In order to obtain true rates that are biologically relevant, characteristic rates were computed based on an empirical data set of the Trypsin protein family (1sgt in table 1), with 83 sequences and 200 sites. The simulation results obtained with the three different vectors of true rates were similar with regard to the relative accuracy of the each model (data not shown). We thus present our results here using the true rates obtained with the *I* model.

True rates, as well as inferred rates, were scaled so that the average was set to 1. For each number of sequences tested, a total of 20 identical and independent simulation runs were conducted. The accuracy of inference was measured using the mean relative absolute deviations (MRAD) distance between the simulated and true rates:

$$\mathrm{MRAD} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\text{estimated}\, r_i - \text{true}\, r_i|}{\text{true}\, r_i}, \tag{13}$$

where M is the sequence length. The division of each absolute deviation by the true rate compensates for the larger variance in large rates and the smaller variance in low rates.

The influence of the number of sequences on the inference accuracy was tested. For this purpose, N sequences ($N = 5$, 10, 15, 20, 25, 30, 35, and 40) were randomly sampled from the original 1sgt data set. Trees of size N were constructed using NJ. A paired t-test was used to determine whether the difference in rate inference was significant between the different models.

Results

Model Comparison

When comparing the fit of the different models to biological data sets, in 82 out of 84 data sets there was a significant improvement in the likelihood under the *D* model as compared with the *I* model. In 60 out of 84 data sets there was a significant improvement in the likelihood under the *D* + *I* model as compared with the *D* model (table 1 contains maximum log-likelihood estimates of ten such data sets, chosen so that the number of sequences in each data set is between 80 and 90).

Twenty of the 24 data sets that did not support the use of the *D* + *I* model shared a common characteristic: these data sets are all very small, where data set size is defined here as the product of the sequence length and the number of sequences. Thus, it may be hypothesized that the addition of the parameters λ_1 and λ_2 is justified only when enough data are available.

Parameter Estimation

The *D* and *D* + *I* models share two common parameters: α, indicative of rate variation, and ρ, which is indicative of the correlation between adjacent sites. When comparing the estimates of these two parameters under the *D* and *D* + *I* models, an interesting pattern emerges. In the vast majority of data sets, α inferred under *D* + *I* is lower than α inferred under *D* ($t = 8.04$, $P < 10^{-11}$) which is subsequently lower than α inferred under *I* ($t = 2.9$, $P < 0.01$). The reverse pattern emerges when analyzing ρ: ρ inferred under *D* + *I* is higher than ρ inferred under *D* ($t = 18.15$, $P < 10^{-30}$). Indeed, in about two-thirds of the data sets, ρ inferred under *D* + *I* is higher than 0.90 (as compared to an average ρ of 0.6 under the *D* model in these data sets), indicating that *D* + *I* managed to differentiate between regions where the rate is highly correlated and regions where this is not so.

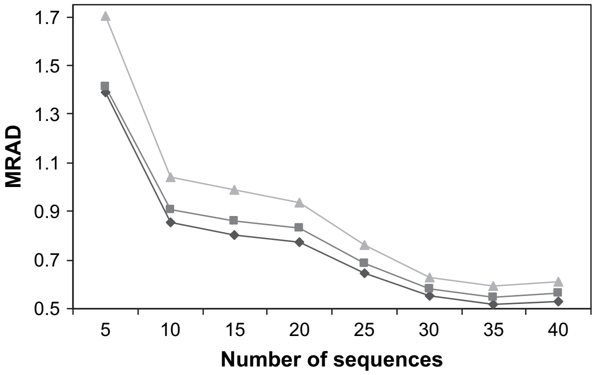


FIG. 4.—Simulation results: accuracy of site-specific rate inference. Diamonds, squares, and triangles represent the $D + I$, D, and I models, respectively.

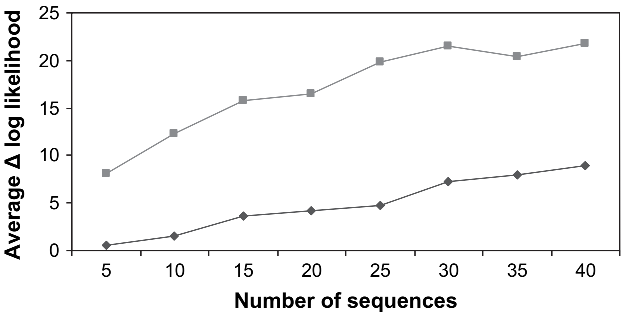


FIG. 5.—Simulation results: average log-likelihood differences between models as a function of number of sequences. Squares represent the difference between the D and the I models, whereas diamonds represent the difference between the $D + I$ and the D models.

Simulation Results

A comparison between the accuracy of site-specific rate inference as a function of the number of sequences under the D and $D + I$ models is shown in figure 4. The difference in accuracy between the D and the $D + I$ models is highly significant ($P < 10^{-4}$), as is the difference between the I and the D models ($P < 0.01$). For a given model, the simulations show that the accuracy increases as the number of sequences increases. This finding is expected because more data are available at each site for rate inference. This accuracy converges when the number of sequences exceeds 30.

Next, we compared the likelihood scores between the different models for each simulation run (fig. 5). On average, the log-likelihood improvement is significant for the D versus the I model for $N \geq 5$ and for $D + I$ versus D for $N \geq 15$. This may be due to the fact that the addition of two parameters in the $D + I$ model cannot be supported by scarce data. In general, in both models it is evident that as the data set size increases, so does the improvement in likelihood.

A Biological Example

Here, we focused on an in-depth analysis of a biological example: K-channels. Such an analysis could point at specific features of the $D + I$ model that are not captured by the I and D models. In addition, it may provide insights as to the relation between the protein structure and function and the predictions of the model.

K-channels function as tetramers and form transmembrane aqueous pores through which K^+ ions can flow. K-channels take part in many different cellular processes including cell volume regulation, hormone secretion, and electrical impulse formation in electrically excitable cells (MacKinnon 2003). The most fundamental role carried out by all K-channels is to allow selective transfer of K^+ ions. The solved 3D structure of a bacterial K^+-channel (Doyle et al. 1998; Jiang et al. 2002) has clarified the mechanism of ion transfer across the membrane.

We used the $D + I$ model to study the K-channel protein family. Fifty-eight homologous sequences (Mayrose, Mitchell, and Pupko 2005) of the channel were used in this study. We focused on two main aspects in studying this protein:

(1) Site-specific rate inference.

(2) Classification of each site to either a dependent rate (S_d) or an independent rate (S_i).

Site-specific rate estimates were projected onto the 3D structure (PDB 1bl8; fig. 6*A*). Two different color-coding schemes were used: one for sites classified as S_d and another for sites classified as S_i. The color code divides the continuous rates inferred into five bins, where bin 1 represents the most variable sites and bin 5 the most conserved sites.

In order to validate the use of the $D + I$ in functional prediction, the results were compared to a previous evolutionary conservation study of the K-channel family (Mayrose, Mitchell, and Pupko 2005). These results were found to be essentially the same as the previous study, matching previous knowledge of important functional regions of the protein. The entrance to the channel, known as the selectivity filter of the channel, is highly conserved (fig. 6*A*). This filter allows only K^+ ions to enter the channel and prevents smaller Na^+ ions from entering (Miller 2000). The inner pore of the channel is lined with hydrophobic residues. These residues are also relatively conserved, and this enables the K^+ ion to travel down an inert pathway.

A valuable feature of the $D + I$ model arises when classifying sites into either S_d or S_i. Excluding a region composed of eight sites which were classified as S_i, all sites of the protein were classified to belong to the S_d category. The rates of these sites classified as S_i sites oscillated between very high and very low values. Examination of the tertiary structure of this region showed that this region is a hydrogen-bonded turn, which is mostly extracellular. Interestingly, this region is uniquely defined by its composition of sites that "zigzag" between being buried and exposed (fig. 6*B*). Moreover, there is a one-to-one correlation between the oscillating pattern of the rates and oscillation in the structure: the sites which are exposed are highly variable, whereas those which are buried are highly conserved. Thus, as previously hypothesized, the region that was classified as belonging to S_i is characterized by a distinctive tertiary structure.

Discussion

To date, the majority of evolutionary models assume that sites evolve independently. Furthermore, even fewer

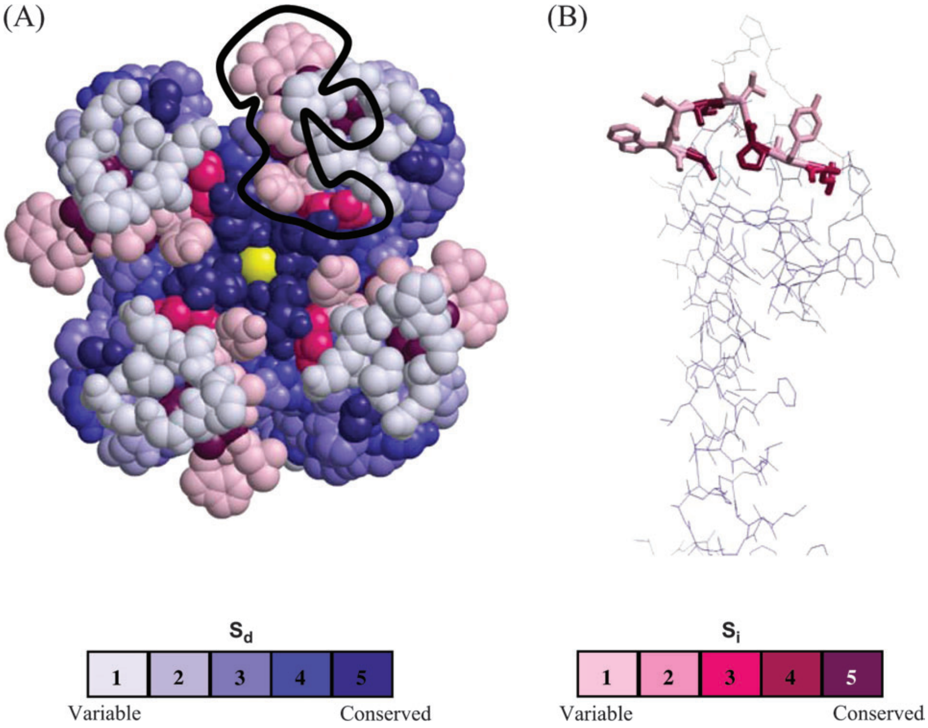


FIG. 6.—The conservation pattern of the K-channel as inferred with $D + I$. Conservation scores together with classification to S_d or S_i are color coded onto the van der Waals surface of the protein. (*A*) The four subunits are viewed from the extracellular side. The circled area represents the region where the sites were classified to S_i. (*B*) One subunit only, viewed from the within the membrane. Sites classified as S_i are marked in thick lines whereas the rest of the molecule is in thin lines. Sites classified as S_i zigzag between exposed (variable; light pink) and buried (conserved; bordeaux).

models take into account the 3D structure of the protein. The model which we have developed here integrates between these two factors, by trying to implicitly incorporate the dependencies between rates in the 3D structure. This is an attempt to reflect the spatial correlation between the evolutionary rates as opposed to a correlation along the linear sequence which is less realistic.

To our knowledge, our presented study is first in comprehensively examining the autocorrelation model (termed *D* here), first presented by Yang (1995). Analysis of a wide variety of data sets shows that taking dependence of adjacent rates into account has a large impact on the likelihood. On average, there was a difference of 74 points between the log likelihood under the *D* model and the log likelihood under the *I* model, with some data sets showing a difference of as many as 500 points. Furthermore, the simulation studies showed an improvement in rate inference accuracy of the *D* model over the classical gamma *I* model. This is especially evident when sequence data are scarce (i.e., $N = 5$). The power of the *D* model lies in its ability to extract more information than the *I* model from an MSA, by extracting spatial information (the dependence between the sites) in addition to the temporal information (the data in a column of the MSA) that both models use. This is an important advantage of the *D* model as compared to other more complex models, which are also more parameter rich. To support such models, large data sets are required. Yet, the *D* model has only two parameters and thus can be supported by less data. To summarize, these results suggest the adequacy of the autocorrelation model, which was mostly neglected by evolutionary studies thus far, for most biological data.

There is a small yet highly consistent improvement of rate inference under the $D + I$ model over the *D* model. Differences in rate estimates can be attributed either to reduced estimation bias or to reduced variance. Our results show that the bias in all three models is the same, and the differences in rate estimation accuracy are a result of reduced variance in the more advanced models (see Supplementary Material online). Furthermore, comparing the rate estimates of the three models on the K-channel data shows that the $D + I$ model creates a smoothing effect in sites classified as S_d and a reverse effect in S_i sites (see Supplementary Material online). The $D + I$ model also shows a consistent improvement of the likelihood throughout the majority of the data sets. It is important to note that the difference in likelihood is merely an underestimate because branch lengths were not optimized in any of the models. The estimation of branch lengths highly affects evolutionary site-specific rate estimation (Pupko et al. 2002). Thus, it is plausible that incorporating branch length estimation into the *D* and $D + I$ models would both improve the likelihood as well as affect the estimation of the rates. However, doing so in linear time is a highly complex task which remains to be resolved.

Our results showed a pattern of correlation between the model parameters α and ρ, where the two parameters seem to compensate one for each other. This pattern can be explained when examining the behavior of the parameters in each model. In the *D* model, the inferred value of ρ represents an average over all sites in the protein. In contrast, the ρ value in the $D + I$ model relates only to those regions where a correlation exists. Regions with rate independence are assigned $\rho = 0$. This leads to the assumption whereby elevated α can compensate for underestimation of ρ. Elevated levels of α represent lower rate variance, which effectively means that most sites share a similar rate. Thus, α contains within itself a measure of dependence between sites, although not necessarily dependence at adjacent sites. This leads to the elevated levels of α in the *D* model, which are meant to compensate for the underestimated ρ, and to the even higher levels of α in the *I* model, where it compensates for ρ being effectively zero.

This is not the first time where α has been found to compensate for another parameter. Sullivan, Swofford, and Naylor (1999) found that underestimation of α leads to underestimation of p_{inv} (a parameter representing the proportion of invariant sites) and vice versa. Similar to Sullivan, Swofford, and Naylor (1999), we also found that the surface of the likelihood function reaches a plateau across the parameter space (data not shown). Thus, estimates of α and ρ as measures of actual properties of the protein should be treated with utmost caution, especially when using simpler models. In such models, it is reasonable to assume that α reflects a combination of several biological properties and not merely the variance of rate variation among sites.

The main strength of the $D + I$ model is revealed in the in-depth study of the K-channel. The study shows that by using the $D + I$ model we may capture unique aspects of a protein 3D structure. This is a novel use of evolutionary rate inference. Currently the main, if not sole, use of evolutionary rates is as predictors of the functional importance of the sites. Here we use evolutionary rates, or rather the relations between the rates, in order to study other characteristics of the protein. It is tempting to consider the use of this model for predicting structure-related features of proteins, namely, features that exhibit a cyclic behavior of the sites—such as beta sheets, alpha helices, and buried-exposed relations. In order to do so, a finer understanding of the distribution of rates throughout the tertiary structure is still required.

Although the $D + I$ and the *D* models show a large improvement in the fit of data sets, a more complex model is still required to explicitly describe the correlation pattern within the 3D structure of proteins. Such a model would have to abandon the classical approach of treating the protein as a linear molecule, and use a general graphical model to describe the relations between the amino acids. In fact, several studies have been published lately which attempt to take into account dependencies of sites along the tertiary structure. One approach involves using a 3D window for detecting selection forces operating on the protein (Suzuki 2004; Berglund et al. 2005). In another approach, the substitution model is constructed so that it takes into account context dependence in the tertiary structure (Robinson et al. 2003; Siepel and Haussler 2004; Rodrigue et al. 2005; Wang and Pollock 2005). It has been shown (Wang and Pollock 2005) that coevolution of sites is highly dependent on the tertiary structure of the protein. Thus, models that account for dependence within the tertiary structure will be able to aid us in the understanding of the mechanism of coevolution. In fact, because protein function is defined by the tertiary structure and by the complex relations between amino acids, it is likely that models which better express this relationship will highly advance our understanding of proteins and their evolution.

Supplementary Materials

Supplementary materials are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank I. Mayrose, A. Doron-Faigenboim, E. Privman, N. Rubinstein, O. Zomer, and two anonymous referees for their insightful comments. T.P. was supported by an Israeli Science Foundation grant number 1208/04 and by a grant in Complexity Science from the Yeshaia Horvitz Association.

Appendix A

Calculation of the likelihood under the *D* model is presented. A similar calculation applies for the $D + I$ model, with the summation extending from 1 through $2K$ states.

$$
\begin{aligned}
L = P(d) &= \sum_{i_1=1}^{K} \cdots \sum_{i_n=1}^{K} P(R_1 = \bar{r}_{i_1}, \ldots, R_n = \bar{r}_{i_n}, d_1, \ldots d_n) \\
&= \sum_{i_1=1}^{K} \cdots \sum_{i_n=1}^{K} P(d_1 \mid d_2, \ldots, d_n, R_1 = \bar{r}_{i_1}, \ldots, R_n = \bar{r}_{i_n}) \\
&\quad \times P(d_2, \ldots, d_n, R_1 = \bar{r}_{i_1}, \ldots, R_n = \bar{r}_{i_n}) \\
&= \sum_{i_1=1}^{K} \cdots \sum_{i_n=1}^{K} P(d_1 \mid R = \bar{r}_{i_1}) \\
&\quad\quad \times P(d_2, \ldots, d_n, R_1 = \bar{r}_{i_1}, \ldots, R_n = \bar{r}_{i_n}) \\
&= \sum_{i_1=1}^{K} \cdots \sum_{i_n=1}^{K} P(d_1 \mid R = \bar{r}_{i_1}) \\
&\quad \times P(d_2 \mid d_3, \ldots, d_n, R_1 = \bar{r}_{i_1}, \ldots, R_n = \bar{r}_{i_n}) \\
&\quad \times P(d_3, \ldots, d_n, R_1 = \bar{r}_{i_1}, \ldots, R_n = \bar{r}_{i_n}) = \ldots \\
&= \sum_{i_1=1}^{K} \cdots \sum_{i_n=1}^{K} P(R_1 = \bar{r}_{i_1}, \ldots, R_n = \bar{r}_{i_n}) \\
&\quad \times \prod_{j=1}^{n} P(d_j | R_j = \bar{r}_{i_j}) \\
&= \sum_{i_1=1}^{K} \cdots \sum_{i_n=1}^{K} P(R_1 = \bar{r}_{i_1}) \\
&\quad \times P(d_1 \mid R_1 = \bar{r}_{i_1}) \cdot \prod_{j=2}^{n} T_{\bar{r}_{i_{j-1}}, \bar{r}_{i_j}} P(d_j \mid R_j = \bar{r}_{i_j}).
\end{aligned}
$$

Literature Cited

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Automatic Control **119**:716–723.

Aloy, P., E. Querol, F. X. Aviles, and M. J. Sternberg. 2001. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J. Mol. Biol. **311**:395–408.

Anisimova, M., J. P. Bielawski, and Z. Yang. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. **18**:1585–1592.

Berglund, A. C., B. Wallner, A. Elofsson, and D. A. Liberles. 2005. Tertiary windowing to detect positive diversifying selection. J. Mol. Evol. **60**:499–504.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. Nucleic Acids Res. **28**:235–242.

Doyle, D. A., J. Morais Cabral, R. A. Pfuetzner, A. Kuo, J. M. Gulbis, S. L. Cohen, B. T. Chait, and R. MacKinnon. 1998. The structure of the potassium channel: molecular basis of K+ conduction and selectivity. Science **280**:69–77.

Durbin, R., S. E. Eddy, A. Krogh, and G. Mitchison. 1998. Biological sequence analysis. Cambridge University Press, Cambridge.

Fares, M. A., S. F. Elena, J. Ortiz, A. Moya, and E. Barrio. 2002. A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. J. Mol. Evol. **55**:509–521.

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

———. 2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies. J. Mol. Evol. **53**:447–455.

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.

Gu, X., Y. X. Fu, and W. H. Li. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. **12**:546–557.

Jiang, Y., A. Lee, J. Chen, M. Cadene, B. T. Chait, and R. MacKinnon. 2002. The open pore conformation of potassium channels. Nature **417**:523–526.

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.

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

MacKinnon, R. 2003. Potassium channels. FEBS Lett. **555**: 62–65.

Mayrose, I., N. Friedman, and T. Pupko. 2005. A gamma mixture model better accounts for among site rate heterogeneity. Bioinformatics Suppl 2: ii151–ii158.

Mayrose, I., D. Graur, N. Ben-Tal, and T. Pupko. 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol. Biol. Evol. **21**:1781–1791.

Mayrose, I., A. Mitchell, and T. Pupko. 2005. Site-specific evolutionary rate inference: taking phylogenetic uncertainty into account. J. Mol. Evol. **60**:345–353.

Miller, C. 2000. An overview of the potassium channel family. Genome Biol. **1**:REVIEWS0004.

Nei, M., and S. Kumar. 2000. Molecular evolution and phylogeny. Oxford University Press, Oxford.

Pupko, T., R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal. 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics **18**(Suppl. 1):S71–S77.

Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. **20**: 1692–1704.

Rodrigue, N., N. Lartillot, D. Bryant, and H. Philippe. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene **347**:207–217.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

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

Sullivan, J., D. Swofford, and G. Naylor. 1999. The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. Mol. Biol. Evol. **16**:1347–1356.

Suzuki, Y. 2004. Three-dimensional window analysis for detecting positive selection at structural regions of proteins. Mol. Biol. Evol. **21**:2352–2359.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. Hillis and B. K. Mable, eds. 2nd ed. Molecular systematics. Sinauer Associates, Sunderland, Mass.

Wang, Z. O., and D. D. Pollock. 2005. Context dependence and coevolution among amino acid residues in proteins. Methods Enzymol. **395**:779–790.

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

———. 1994. 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.

———. 1996. Among-site variation and its impact on phylogenetic analyses. Trends Ecol. Evol. **11**:367–372.

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

Martin Embley, Associate Editor

Accepted October 19, 2005