BIOMETRICS 51, 552–561
June 1995

Mixed Model Analysis of DNA Sequence Evolution

Ziheng Yang[1] and Tianlin Wang[2]

[1]Department of Zoology, University of Cambridge,
Downing Street, Cambridge CB2 3EJ, United Kingdom
[2]Department of Animal Sciences, University of Illinois,
1207 West Gregory Drive, Urbana, Illinois 61801, U.S.A.

SUMMARY

Nucleotides in a DNA sequence may be changing at different rates, because they are located in different structural and functional regions of the gene, and are thus subject to different mutational pressures or selective restrictions. Knowledge of substitution rates at specific sites is important for understanding the forces and mechanisms that have shaped the evolution of the DNA sequences. The gamma distribution has previously been proposed to model such rate variation among nucleotide sites. Based on mixed model methodology we present in this paper a method for predicting substitution rates at nucleotide sites by using homologous DNA sequences. The predictor is unbiased and "best" in the sense that it minimizes the mean squared error and maximizes the correlation between the predictor and the true value. It is also quite robust to errors in estimates of parameters in the model. A numerical example is given, with guidelines for the practical use of the approach. The most influential factor affecting the accuracy of prediction is the number of sequences; to get a correlation of over .7 between the predictor and the true value, about six to seven sequences are needed, depending on the overall similarity of the sequences.

1. Introduction

The deoxyribonucleic acid (DNA) sequences of living organisms provide important information on the relationships among the species and on the evolutionary process giving rise to the sequences. By comparing homologous sequences across species, we can not only make inferences concerning major evolutionary events, but also approach a better understanding of the forces and mechanisms of molecular evolution. Felsenstein (1981) presented a maximum likelihood framework for modeling the evolution of DNA sequences that are linked by a tree structure. The method has been widely used to estimate evolutionary trees from DNA sequence data, and appears superior to other methods principally based on intuitive arguments (e.g., Hasegawa, Kishino, and Saitou, 1991; Yang, 1994).

Probably the most worrying assumption made in the model of Felsenstein (1981) is that substitution rates are constant across nucleotide sites, which is unrealistic at least for sequences with biological functions. Wakeley (1993) provided an excellent review of evidences of rate variation over sites and of early studies on this problem. Recently Yang (1993) suggested an extension to the model of Felsenstein (1981), in which a gamma distribution is used to describe such spatial rate variation. In this formulation, rates at different sites are taken to be random variables from the gamma distribution and are integrated out in the likelihood function. The new model was found to fit real data rather well (Yang, Goldman, and Friday, 1994).

However, knowledge of the rates themselves is also very useful. First, it is well-known that different sites or regions in a gene are liable to be changing at different rates, because they are of different significance to the structure and function of the gene or protein, and thus under different selective restrictions. Even mutation rates are believed to vary at different regions of the genome. Predicting those rates may help us understand the mechanisms of molecular sequence evolution. Second, it has been suspected that when the rates are variable over sites, rates at neighboring sites

Key words: Best unbiased predictor; DNA sequences; Empirical Bayes estimation; Gamma distribution; Maximum likelihood method; Mixed models; Nucleotide substitution; Spatial rate variation.

may be (positively) correlated because of the existence of "variable" and "conservative" regions in a gene. By studying the pattern in real sequences, we can gain some insights into the process, which could help the formulation of a model that can allow such auto-correlated rates.

In Section 2 of this paper we propose a method for predicting substitution rates at sites, based on mixed model methodology. The predictor is the conditional mean of the rate given the observed data. In Section 3 we will examine the properties of the predictor and the effects of parameter values on the accuracy of prediction. In Section 4 the method is applied to data comprising the mitochondrial DNA (mtDNA) sequences from several primate species. Special attention is directed to the robustness of the predictor to errors in estimates of parameters in the model. Finally in Section 5 we discuss the potential uses of the approach.

2. Theory

2.1 *The Model*

2.1.1 *The data.* Let the number of species (sequences) be s and the length of sequence be n. The data can be denoted by an $s \times n$ matrix $\mathbf{X} = \{x_{ij}\}$, where x_{ij} means the jth nucleotide in the ith sequence. x_{ij} takes a value from 1, 2, 3, or 4, representing the four nucleotides, T, C, A, or G, respectively. We assume that nucleotide substitutions at different sites are independent and follow the same general stochastic process. Therefore one datum corresponds to one column in $\mathbf{X}$, such as $\mathbf{x}_j$. The $\mathbf{x}_j$ ($j = 1, 2, \ldots, n$) are then identically and independently distributed, and follow a multinomial distribution. The number of categories in the distribution, that is, the number of possible site patterns, is 4^s, with the probability of occurrence in each category determined by the tree and the parameters in the model.

2.1.2 *The tree.* The sequences are assumed to be related by a tree structure. The single unrooted tree, T, for three species, is shown in Figure 1. The branch lengths, $\mathbf{t} = \{t_1, t_2, t_3\}$, are measured by the average numbers of nucleotide substitutions per site along the branches, to be explained below.

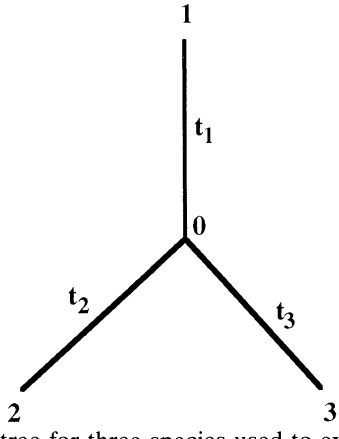

Figure 1. The unrooted tree for three species used to explain the prediction method

2.1.3 *The pattern of nucleotide substitution.* Nucleotide substitution is assumed to follow a stationary homogeneous Markov process, with the four nucleotides to be the states of the process. Let $\mathbf{Q} = \{Q_{ij}\}$ be the 4×4 rate matrix of the process for a site with an average overall rate. In this study, we use the model of Hasegawa, Kishino, and Yano (1985), designated "HKY85," by which the probability that a given nucleotide, i, will change into nucleotide j ($j \neq i$), in a small time interval Δt, is given by

$$Q_{ij}\Delta t = \begin{cases} \kappa \mu \pi_j \Delta t \text{ (for transitions: T}\leftrightarrow\text{C, A}\leftrightarrow\text{G)} \\ \mu \pi_j \Delta t \text{ (for transversions: T}\leftrightarrow\text{A, T}\leftrightarrow\text{G, C}\leftrightarrow\text{A, C}\leftrightarrow\text{G)} \end{cases} \quad (1)$$

where π_j is the equilibrium frequency of nucleotide j, with $\Sigma\, \pi_j = 1$. The row sums of $\mathbf{Q}$ are zero and therefore the diagonals are given by $Q_{ii} = -\Sigma_{j \neq i} Q_{ij}$ (Grimmett and Stirzaker, 1992). As it is not possible to estimate time and rate separately, we choose the scale factor μ such that the expected rate of substitution is 1, that is, $-\Sigma\, \pi_i Q_{ii} = 1$. Time t or branch lengths in a tree will then be measured by the distance, that is, the expected number of nucleotide substitutions per site.

To calculate the matrix of transition probabilities, $\mathbf{P}(t) = \exp(\mathbf{Q}t)$, we perform the spectral decomposition (diagonalization) of $\mathbf{Q}$, that is, $\mathbf{Q} = \mathbf{U}\, \mathrm{diag}\{\lambda_0, \lambda_1, \lambda_2, \lambda_3\}\, \mathbf{V}$, where λ_0, λ_1, λ_2, and λ_3

are the eigenvalues of **Q**, columns of **U** are the corresponding right eigenvectors, and rows of **V** are the corresponding left eigenvectors, with $\mathbf{U} = \mathbf{V}^{-1}$. The full eigensolution of **Q** for the HKY85 model has been given by Hasegawa et al. (1985). Then we have

$$\mathbf{P}(t) = \mathbf{U}\ \mathrm{diag}\{\exp(\lambda_0 t),\ \exp(\lambda_1 t),\ \exp(\lambda_2 t),\ \exp(\lambda_3 t)\}\mathbf{V}. \tag{2}$$

The ijth element of $\mathbf{P}(t)$ is given by

$$P_{ij}(t) = \sum_{k=0}^{3} U_{ik}V_{kj}\exp(\lambda_k t) = \sum_{k=0}^{3} c_{ijk}\exp(\lambda_k t), \tag{3}$$

where U_{ik} and V_{kj} are elements of matrices **U** and **V**, respectively, and $c_{ijk} = U_{ik}V_{kj}$.

2.1.4 *Rate variation over nucleotide sites.* For any given site, we assume that the rate matrix is $r\mathbf{Q}$, where the overall rate r is a random variable from a gamma distribution. The gamma distribution with parameters α and β has mean α/β and variance α/β^2. We set $\beta = \alpha$ so that $\mathrm{E}(r) = 1$ and then $\mathrm{var}(r) = 1/\alpha$. The probability density function of r is therefore

$$f(r) = \alpha^{\alpha}\Gamma(\alpha)^{-1}e^{-\alpha r}r^{\alpha-1},\ r > 0,\ \alpha > 0. \tag{4}$$

The tree structure, T, the branch lengths **t** of the tree, the rate and frequency parameters in the HKY85 model, κ, π_T, π_C, π_A, and the variance parameter of the gamma distribution, α, are all parameters in the model, collectively denoted $\boldsymbol{\theta} = \{T, \mathbf{t}, \kappa, \pi_T, \pi_C, \pi_A, \alpha\}$. In the terms of mixed model theory, $\boldsymbol{\theta}$ may be called "fixed effects" while the r values are "random effects." Yang (1993) presented a procedure for the maximum likelihood estimation of $\boldsymbol{\theta}$, while our objective in the present paper is to predict r for a given datum **x**.

2.2 *The Prediction Method*

Based on the mixed model theory (e.g., Henderson, 1973), we study $f(r|\mathbf{x};\ \boldsymbol{\theta})$, the distribution of r given **x**. The point predictor, $\hat{r}$, can be defined as the conditional expectation

$$\hat{r} = \mathrm{E}(r|\mathbf{x};\ \boldsymbol{\theta}) = \int_0^{\infty} rf(r|\mathbf{x};\ \boldsymbol{\theta})dr$$

$$= \int_0^{\infty} rf(\mathbf{x}|r;\ \boldsymbol{\theta})f(r)dr/f(\mathbf{x};\ \boldsymbol{\theta}) \tag{5}$$

Consider the three species problem of Figure 1. Let $\mathbf{x} = \{x_1, x_2, x_3\}'$ be a column in the data matrix **X**. As our model cannot identify the root of the tree (Felsenstein 1981), we can arbitrarily set the "root" at node 0 in Figure 1. It then follows that (Yang, 1993)

$$f(\mathbf{x}|r;\ \boldsymbol{\theta}) = \sum_{x_0=1}^{4} \pi_{x_0}[P_{x_0x_1}(t_1 r)P_{x_0x_2}(t_2 r)P_{x_0x_3}(t_3 r)] \tag{6}$$

$$= \sum_{x_0=1}^{4} \pi_{x_0} \sum_{m=1}^{4^3} [c_{x_0x_1M_{m1}}c_{x_0x_2M_{m2}}c_{x_0x_3M_{m3}}] \exp\left(r\sum_{j=1}^{3}\lambda_{M_{mj}}t_j\right) \tag{7}$$

$$= \sum_{m=1}^{4^3} B(m, \mathbf{x})\exp(rS_m), \tag{8}$$

where $B(m, \mathbf{x}) = \sum_{x_0=1}^{4}\pi_{x_0}[c_{x_0M_{m1}}c_{x_0x_2M_{m2}}c_{x_0x_3M_{m3}}]$ and $S_m = \sum_{j=1}^{3}\lambda_{M_{mj}}t_j$. The product in the square bracket in (6) is expanded to give (7), where M_{mj} is the jth digit when $(m-1)$ is expressed as a number of base 4.

It follows from (4) and (8) that

$$\int rf(\mathbf{x}|r;\ \boldsymbol{\theta})f(r)dr = \sum_{m=1}^{4^3} B(m, \mathbf{x})\left[\frac{\alpha}{\alpha - S_m}\right]^{\alpha+1}, \tag{9}$$

while $f(\mathbf{x};\ \theta)$ is given by Yang (1993) as

$$f(\mathbf{x};\ \boldsymbol{\theta}) = \int f(\mathbf{x}|r;\ \boldsymbol{\theta})f(r)dr = \sum_{m=1}^{4^3} B(m,\ \mathbf{x})\left[\frac{\alpha}{\alpha - S_m}\right]^{\alpha}. \tag{10}$$

Substituting (9) and (10) into (5), we have

$$\hat{r} = \frac{\sum_{m=1}^{4^3} B(m,\ \mathbf{x})\left[\dfrac{\alpha}{\alpha - S_m}\right]^{\alpha+1}}{\sum_{m=1}^{4^3} B(m,\ \mathbf{x})\left[\dfrac{\alpha}{\alpha - S_m}\right]^{\alpha}}. \tag{11}$$

In practice, $\boldsymbol{\theta}$ will be replaced by their estimates. In this paper, we estimate $\boldsymbol{\theta}$ from the likelihood function based on $f(\mathbf{x};\ \boldsymbol{\theta})$ (Yang, 1993). This is justified by the invariance property of maximum likelihood estimators (MLE)

$$\text{MLE}\{\text{E}(r|\mathbf{x};\ \boldsymbol{\theta})\} = \text{E}\{r|\mathbf{x};\ \text{MLE}(\boldsymbol{\theta})\}. \tag{12}$$

From a Bayesian point of view, r is a parameter, with a gamma prior distribution $f(r)$. $f(r|\mathbf{x})$ is then the posterior distribution. Use of the square error loss will lead to $\hat{r}$ (5) as the Bayesian point estimator of r (Maritz and Lwin, 1989). Replacement of $\boldsymbol{\theta}$ by their estimates will have an empirical Bayesian justification (Maritz and Lwin, 1989). The distribution $f(r|\mathbf{x})$ can also be used to construct the credibility set of the predictor.

3. Evaluation of the Prediction Method

3.1 *Properties of the Predictor*

By examining the prediction error, $\hat{r} - r$, we note that $\hat{r}$

(1) is unbiased, $\text{E}(\hat{r} - r) = 0$;
(2) has the smallest mean squared error, $\text{MSE}(\hat{r}) = \text{E}(\hat{r} - r)^2$;
(3) has the highest correlation with r, $\rho(\hat{r}, r)$.

As the predictor is unbiased, the MSE is equal to the variance of its prediction error. Thus the second property can also be stated as $\hat{r}$ having the smallest prediction error variance.

Let $\bar{r} = g(\mathbf{x})$ be another predictor of r, with $g(\mathbf{x})$ to be any function of $\mathbf{x}$. It follows that

$$\begin{aligned}\text{E}(\bar{r}r) &= \int\int \bar{r}rf(\mathbf{x}|r)f(r)dr\,d\mathbf{x} \\ &= \int \bar{r}\hat{r}f(\mathbf{x})\,d\mathbf{x} \\ &= \text{E}(\bar{r}\hat{r})\end{aligned} \tag{13}$$

where $\int d\mathbf{x}$ means summation over $\mathbf{x}$ as $\mathbf{x}$ is discrete. Similarly we have $\text{E}\{(\bar{r} - \hat{r})r\} = \text{E}\{(\bar{r} - \hat{r})\hat{r}\}$. Therefore

$$\begin{aligned}\text{E}(\bar{r} - r)^2 &= \text{E}(\bar{r} - \hat{r})^2 + \text{E}(\hat{r} - r)^2 + 2\text{E}\{(\bar{r} - \hat{r})(\hat{r} - r)\} \\ &= \text{E}(\bar{r} - \hat{r})^2 + \text{E}(\hat{r} - r)^2,\end{aligned} \tag{14}$$

which proves the second property.

Furthermore, we note that $\rho^2(\hat{r}, r) = \text{var}(\hat{r})/\text{var}(r)$. Equation (13) implies $\text{cov}(\bar{r}, r) = \text{cov}(\bar{r}, \hat{r})$, which leads to the following equation, proving the third property

$$\rho^2(\bar{r}, r) = \rho^2(\bar{r}, \hat{r}) \cdot \rho^2(\hat{r}, r). \tag{15}$$

3.2 *The Accuracy of Prediction as a Function of Parameters*

We assess the accuracy of prediction by $\rho(\hat{r}, r) = [\text{var}(\hat{r})/\text{var}(r)]^{1/2}$, with $\boldsymbol{\theta}$ replaced by the maximum likelihood estimates. In this subsection we examine the effects of the branch lengths $\mathbf{t}$, and parameters κ and α. The effects of the number of sequences, and of the tree topology, and the robustness of the prediction method to possible errors in estimates of parameters will be examined with real data in the next section.

Due to the complexity of the parameter space, we study only some special cases. The example tree to be used is shown in Figure 2, where a measures the overall amount of evolution. The HKY85 model of nucleotide substitution is adopted, with frequency parameters fixed at $\pi_T = .1$, $\pi_C = .2$, $\pi_A = .3$ and $\pi_G = .4$.

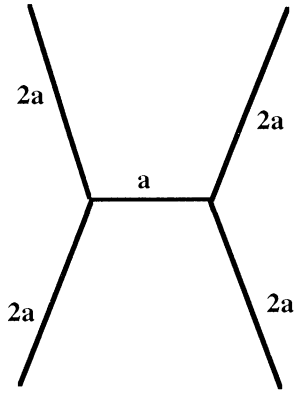

Figure 2. An example tree of four species used for evaluating the accuracy of prediction

Figure 3 shows the effect of branch lengths, as measured by a. We assume $\alpha = .25$ and $\kappa = 10$. When branch lengths are either very small or very large, the sequences will be either very similar or very different and the data will contain little evolutionary information. We can therefore expect the highest accuracy to occur with medium branch lengths, as show by $\rho(\hat{r}, r)$ and MSE($\hat{r}$) in Figure 3.

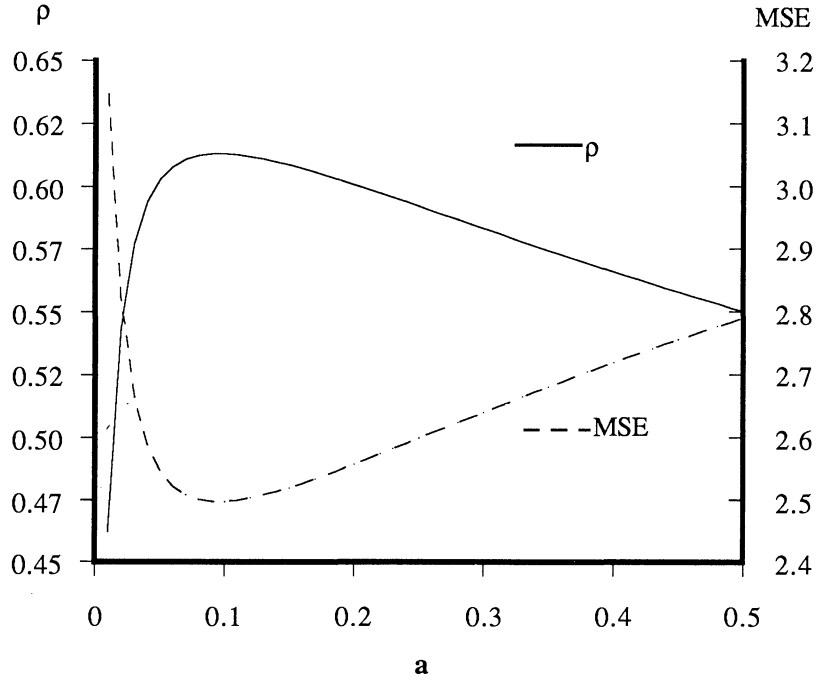

Figure 3. The accuracy of prediction, as measured by $\rho(\hat{r}, r)$ and MSE($\hat{r}$), as a function of a in Figure 2

Figure 4 shows the combined effect of a and α. Parameter κ is fixed at 10. Obviously α can influence ρ substantially. The prediction is very poor when α is very large, that is, when there is little rate variation over sites. Estimates of α from real data are normally in the range .1–1 (Yang et al.

(1994) and unpublished results), in which cases ρ is high. For given α, ρ is highest with medium values of a, which is consistent with Figure 3.

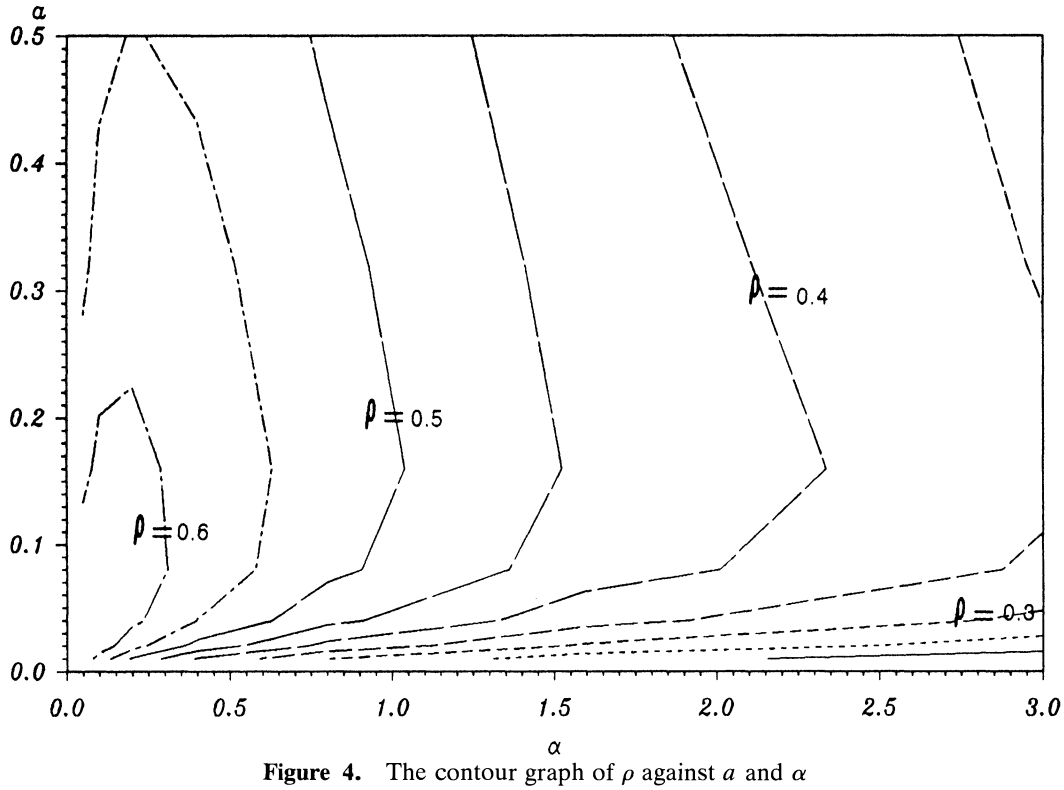

Figure 4. The contour graph of ρ against a and α

The effects of κ and α are shown in Figure 5. a is fixed at 0.1. Once again, α is an important factor. The κ parameter, when estimated from real data, is usually in the range 1–10 (Yang et al., 1994) and unpublished results), and thus has very little effect on ρ.

4. Analysis of the mtDNA Sequences From Several Primate Species

4.1 *The Predicted Substitution Rates*

Brown et al. (1982) determined the sequences of a segment of mitochondrial genomes of human, chimpanzee, gorilla, orangutan, and gibbon. The aligned sequences contain 895 nucleotide sites. For ease of computation we will only use the human, chimpanzee, gorilla, and orangutan sequences.

The HKY85 substitution model is used, with gamma-distributed rates over sites (HKY85+Γ). The frequency parameters in the HKY85 model are estimated by averaging the observed values over species: $\hat{\pi}_T = .2542$, $\hat{\pi}_C = .3313$, $\hat{\pi}_A = .3106$, $\hat{\pi}_G = .1039$. The other parameters are estimated by maximizing the likelihood function; $\hat{\kappa} = 30.05 \pm 14.67$ and $\hat{\alpha} = .20 \pm .07$ (standard errors are estimated by the curvature method). The unrooted maximum likelihood tree separates the species in the order human, chimpanzee, gorilla, and orangutan; that is, the tree topology can be represented as ((human, chimpanzee), gorilla, orangutan).

In Figure 6 the predicted rate is plotted with the expected number of occurrences of the site pattern, which is virtually the same as the observed number. The graph looks like a gamma distribution. The correlation calculated from the estimated values of parameters is $\rho = .592$. Because we replaced the parameters with their estimates, this should be taken as the upper limit of the accuracy of prediction. Our results below suggest, however, that this upper limit is, indeed, almost reached in practice.

It is noteworthy that Figure 6 involves no "sampling errors" caused by limited data, as $f(\mathbf{x})$ is the expected frequency for site pattern $\mathbf{x}$. The predicted rates involve errors mainly because our model only provides one $\hat{r}$ for each $\mathbf{x}$, although a datum $\mathbf{x}$ can by generated by different (real) rates. The accuracy of prediction cannot be improved by increasing the number of data n, although this will lead to more reliable estimates of $\boldsymbol{\theta}$.

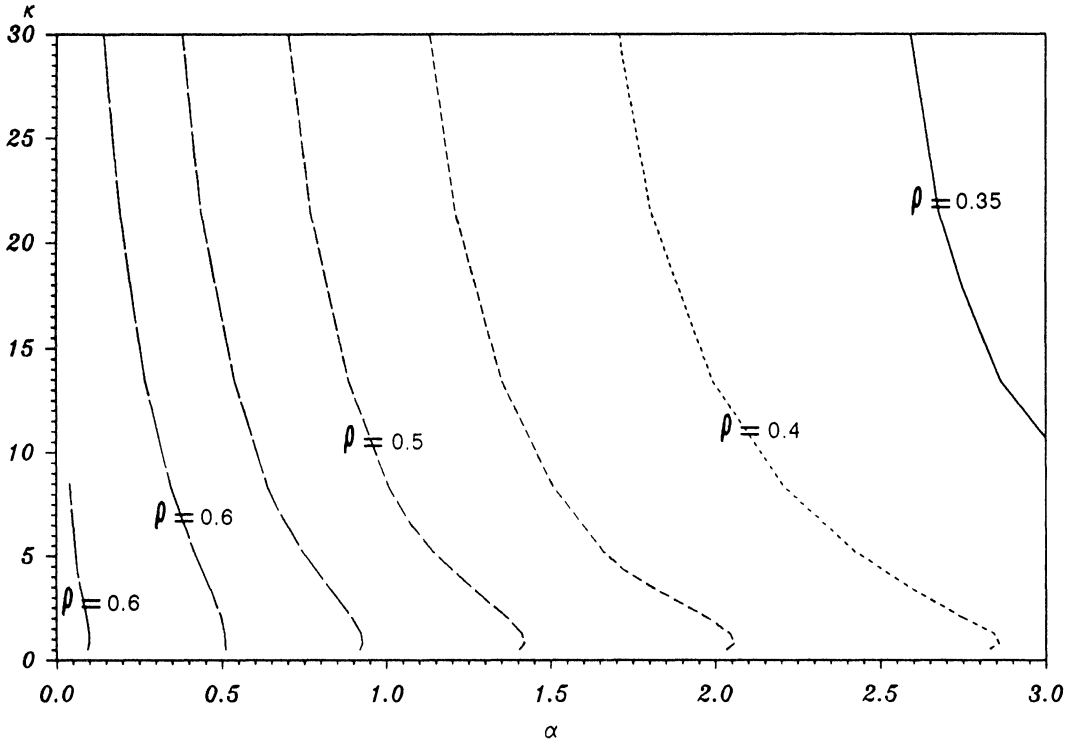

Figure 5. The contour graph of ρ against κ and α

4.2 *The Robustness of the Prediction Method to Errors in Parameter Estimates*

With four species, there are three unrooted bifurcating trees and one star-like tree. In a previous study, we have found that estimates of parameters in the evolutionary model, for instance, those of κ and α, are very stable across different tree topologies (Yang et al., 1994). In this study, we find that predicted rates are also stable no matter which tree is assumed. For instance, the correlation, calculated from the observed site pattern frequencies, between the rates obtained by using the maximum likelihood tree and those obtained by using the star-like tree, is .985. By (15), it is obvious that the reduction in accuracy of prediction by using a wrong tree, such as the star-like tree, is trivial (1–2%).

The predicted values of rates are also very insensitive to changes in parameter estimates. The estimates of κ and α are found to be negatively correlated (Yang et al., 1994). When κ is fixed at 12, the estimate when a single rate over sites is assumed, $\hat{\alpha} = .458$. Branch lengths are then seriously underestimated. Nevertheless, the rates predicted by using these values of κ and α have a correlation of .981 with those obtained by using the real MLEs of the parameters; therefore the twofold difference in these parameters has not caused substantial reduction in ρ. Similarly, it can be expected that using the MLEs of the parameters instead of their (unknown) true values will not cause considerable reduction in ρ.

4.3 *The Number of Sequences*

The mtDNA sequence data have been expanded since the publication of Brown et al. (1982). In the following we analyze a larger dataset containing homologous sequences of this same region from human, chimpanzee, gorilla, orangutan, gibbon, crab-eating macaque, tarsier, and lemur. The sequences are aligned by Adrian Friday. After sites involving insertions and deletions are excluded, there are $n = 888$ nucleotides in the sequence.

The sequences were added into the dataset one by one, beginning from three. Parameter κ was fixed at 10 for all the analyses and only the star trees are evaluated. The α parameter and branch lengths in the star trees were obtained by iteration. Estimates of α are .445, .436, .317, .336, .335, .316, and .283 when $s = 3, 4, \ldots, 9$, respectively. ρ was calculated using two methods. The "exact" method sums over all the possible data outcomes, while the "approximate" method makes use of the observed site pattern frequencies. As expected, the improvement in ρ gets smaller and smaller when more and more sequences are added to the dataset (Figure 7).

It should be noted that ρ is somewhat overestimated. First, we note that for the same data, the ρ

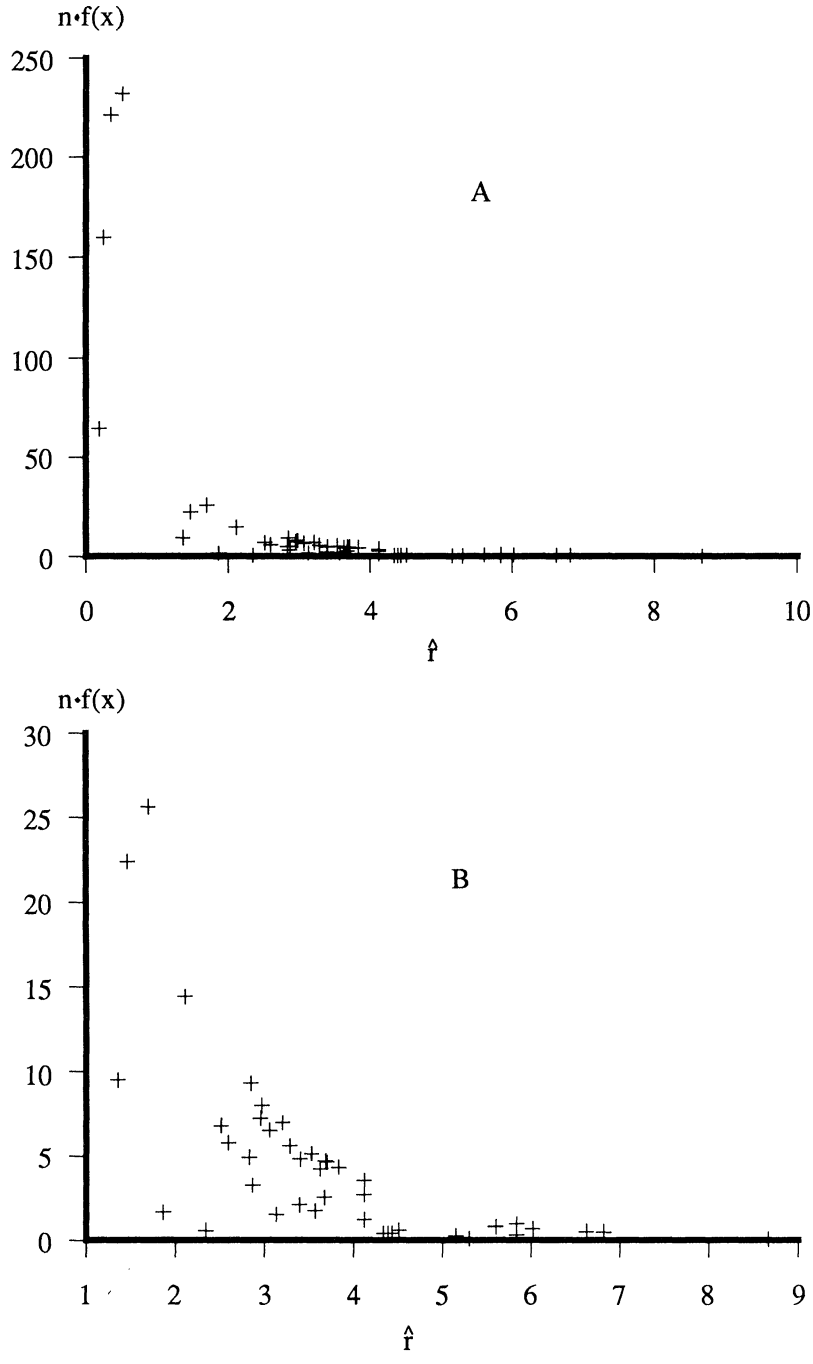

Figure 6. The predicted rate for each site pattern plotted with the expected number of occurrances of the site pattern. The 895-base pair mtDNA data for four species are analyzed using the HKY85+Γ model. Only 46 out of the $4^4 = 256$ possible site patterns are actually observed in the data and only they are shown in the graph. One site pattern corresponds to one point in the graph. The four most frequent site patterns are AAAA, TTTT, CCCC, and GGGG, and predicted rates for those four site patterns only are less than 1, the average. *A*, all the 46 observed site patterns; *B*, 42 site patterns, with the four most frequent site patterns excluded.

calculated by using the star tree is slightly higher than that obtained by using a bifurcating tree. For example, for the mtDNA data for four species, analyzed in the previous subsection, $\rho = .592$ by the maximum likelihood tree, while $\rho = .600$ by the star tree. This discrepancy is expected to be larger when more sequences are analyzed. In other words, the increase in ρ is slower than is suggested by

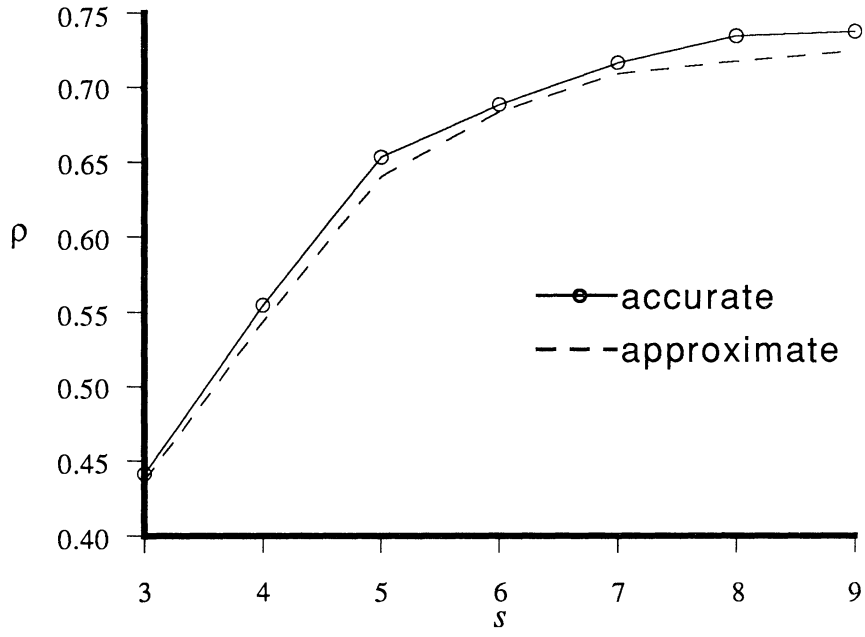

Figure 7. The accuracy of prediction, ρ, as a function of the number of species in the dataset. The nine mtDNA sequences of human, chimpanzee, gorilla, orangutan, gibbon, macaque, tarsier, and lemur are added into the dataset in this order. Only the star trees are evaluated.

Figure 7. Second, as the star trees instead of the true trees are assumed, parameter estimates used for predicting the rates are neither the true values nor their genuine maximum likelihood estimates. Substitution rates predicted this way will involve biases. This effect appears to be minor, however, judging from the results of the previous section.

5. Discussion

The major limitation on the practical use of the prediction method proposed in this paper is the maximum likelihood estimation of $\boldsymbol{\theta}$, especially the tree structure, T. Calculation of $f(\mathbf{x}; \boldsymbol{\theta})$ by the method of Yang (1993) requires intensive computation. Currently only four or five species can be handled in an endurable amount of computation time if all the tree topologies are to be compared. The prediction of substitution rates, however, needs much less computation, equivalent to one round of evaluation of the likelihood function. Exact calculation of ρ involves much more computation when $s > 5$.

One solution that appears acceptable is to use only the star tree to estimate parameters and predict rates. When only the star tree is evaluated, many more sequences can be handled by the algorithm. We note that the prediction of substitution rates is tolerant to errors in estimates of the tree and other parameters. ρ appear to be more sensitive to such errors, but as the estimates of parameters are quite stable over tree topologies (Yang et al., 1994), even ρ is reliably estimated.

Perhaps the most serious unrealistic assumption made in the present model is the independence of nucleotide substitutions among nucleotide sites. In fact one of our objectives in predicting rates is to study their possible auto-correlations. Some element of contradiction is thus involved. At the moment, even a simple model of first-order autocorrelation of rates does not appear to be computationally feasible. However, we suggest that the maximum likelihood estimation of $\boldsymbol{\theta}$ is not seriously affected by the violation of the independence assumption. The prediction of rates may be influenced to a greater extent, but we believe such errors to be acceptably small.

The stability, across tree topologies, of predicted rates as observed in this paper, and of parameter estimates as observed by Yang et al. (1994), may have practical implications concerning tree reconstruction. Specifically, parameters can be estimated from the star tree and later be used in the evaluation of other tree topologies. The rates predicted using the star tree may be used to collapse the sites into, say, five or ten classes, with one average rate for all the sites within each class. Such rates for classes of sites can then be used to compare the other tree topologies. The computation involved with this approximate method will then be comparable to that of Felsenstein's (1981) method which assumes a single rate over sites, and its performance relative to the exact method of Yang (1993) appears to be an interesting open question.

ACKNOWLEDGEMENTS

We are very grateful to Adrian Friday for many constructive comments on an earlier version of the manuscript, and for preparing and aligning the nine-species mtDNA sequences analyzed in this paper. Z. Y. thanks Adrian Friday and Nick Goldman for discussions during this study. T. W. wishes to thank the Department of Animal Sciences, University of Illinois for financial support.

RÉSUMÉ

Les nucléotides dans une séquence d'ADN peuvent changer à différents taux parce que situés dans différentes régions structurelles et fonctionnelles du gène; ils sont donc sujets à des pressions de mutation ou à des restrictions sélectives différentes. La connaissance des taux de substitution en des sites spécifiques est importante pour comprendre les forces et mécanismes qui ont modelé l'évolution des séquences d'ADN. La distribution gamma a déjà été proposée pour modeliser un tel taux de variation parmi les sites nucléotidiques. Dans ce papier fondé sur la méthodologie du modèle mixte, nous présentons une méthode de prédiction des taux de substitution aux sites nucléotidiques en utilisant des séquences d'ADN homologues. Le prédicteur est non biaisé et le "meilleur" au sens où il minimise l'erreur quadratique moyenne et maximise la corrélation entre le prédicteur et la vraie valeur. Il est aussi plutôt robuste aux erreurs dans les estimations des paramètres du modèle. On donne un exemple numérique, avec des recommendations pour l'usage pratique de cette approche. Le facteur le plus influent de la précision de prédiction est le nombre de séquences; il est nécessaire de disposer de six ou sept séquences pour obtenir une corrélation d'au moins .7 entre la prédiction et la vraie valeur; ceci dépend de la similarité globale des séquences.

REFERENCES

Brown, W. M., Prager, E. M., Wang, A., and Wilson, A. C. (1982). Mitochondrial DNA sequences of primates, tempo and mode of evolution. *Journal of Molecular Evolution* **18,** 225–239.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17,** 368–376.

Grimmett, G. R. and Stirzaker, D. R. (1992). *Probability and Random Processes*, 2nd Ed., pp. 239–246. Oxford: Clarendon Press.

Hasegawa, M., Kishino, H., and Saitou, N. (1991). On the maximum likelihood method in molecular phylogenetics. *Journal of Molecular Evolution* **32,** 443–445.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22,** 160–174.

Henderson, R. (1973). Sire evaluation and genetic trends. In *Animal Breeding and Genetics Symposium in Honour of Dr. J. L. Lush*, pp. 10–41. Champaign, Illinois: American Society of Animal Science and Animal Dairy Science Association.

Maritz, J. S. and Lwin, T. (1989). *Empirical Bayes Methods*. London: Chapman and Hall.

Wakeley, J. (1993). Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. *Journal of Molecular Evolution* **37,** 613–623.

Yang, Z. (1993). Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* **10,** 1396–1401.

Yang, Z. (1994). Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Systematic Biology* ***43,*** 329–342.

Yang, Z., Goldman, N., and Friday, A. E. (1994). Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. *Molecular Biology and Evolution*. ***11,*** 316–324.

Received September 1993; *revised January* 1994; *accepted February* 1994.