A Likelihood Approach for Comparing Synonymous and Nonsynonymous Nucleotide Substitution Rates, with Application to the Chloroplast Genome

Spencer V. Muse[1] *and Brandon S. Gaut*
Program in Statistical Genetics, Department of Statistics, North Carolina State University

A model of DNA sequence evolution applicable to coding regions is presented. This represents the first evolutionary model that accounts for dependencies among nucleotides within a codon. The model uses the codon, as opposed to the nucleotide, as the unit of evolution, and is parameterized in terms of synonymous and nonsynonymous nucleotide substitution rates. One of the model's advantages over those used in methods for estimating synonymous and nonsynonymous substitution rates is that it completely corrects for multiple hits at a codon, rather than taking a parsimony approach and considering only pathways of minimum change between homologous codons. Likelihood-ratio versions of the relative-rate test are constructed and applied to data from the complete chloroplast DNA sequences of *Oryza sativa, Nicotiana tabacum,* and *Marchantia polymorpha.* Results of these tests confirm previous findings that substitution rates in the chloroplast genome are subject to both lineage-specific and locus-specific effects. Additionally, the new tests suggest that the rate heterogeneity is due primarily to differences in nonsynonymous substitution rates. Simulations help confirm previous suggestions that silent sites are saturated, leaving no evidence of heterogeneity in synonymous substitution rates.

Introduction

An understanding of nucleotide substitution rates is of fundamental importance in the field of molecular evolution, and a great deal of progress has been made in this area. Partitioning the total substitution rate into synonymous and nonsynonymous components is one of the primary objectives of evolutionary studies involving coding regions. Unfortunately, estimation of these parameters is not straightforward. Several useful methods have been presented, but all of them rely on a conceptual framework that is not appropriate in some cases. The goal of this work is to present a generally applicable model of DNA coding-sequence evolution that is parameterized in terms of synonymous and nonsynonymous nucleotide substitution rates and to demonstrate some of its potential applications. New likelihood methods for performing relative-rate tests are developed, and these tests are applied to data from the complete chloroplast DNA (cpDNA) sequences of three highly diverged plant species. It is shown that these tests allow a more precise description of the substitution process than did previous methods.

Key words: relative-rate test, codon evolution, evolutionary model, cpDNA, likelihood ratio.

1. Present address and address for correspondence and reprints: Spencer V. Muse, Institute of Molecular Evolutionary Genetics, Department of Biology, 208 Mueller Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802-5301.

Mol. Biol. Evol. 11(5):715–724. 1994.

0737-4038/94/1105-0001$02.00

Methods

Existing Methods

Existing methods for performing relative-rate tests on synonymous and nonsynonymous substitution rates are based on methods for estimating these parameters. Several such methods exist. Miyata and Yasunaga (1980), Perler et al. (1980), and Li et al. (1985*b*) each proposed methods that, to varying degrees, take into consideration the dependencies between nucleotides within a codon and the lack of symmetry in the genetic code. Both Li et al. (1985*a*) and Nei and Gojobori (1986) review this literature, and in the latter a simplified version of the method of Miyata and Yasunaga (1980) is proposed as a satisfactory alternative.

The procedures mentioned above rest on classifying sites into "degeneracy classes." At fourfold-degenerate sites no substitutions cause amino acid substitutions. That is, any one of the four nucleotides may occupy these positions without altering the encoded amino acid. The third positions of 32 of the 61 nontermination codons fall into this class. At twofold-degenerate sites (with minor exceptions) transitions are synonymous, while transversions are nonsynonymous. There are 24 third-position sites and 8 first-position sites in this category. All changes at nondegenerate sites lead to amino acid

	T		C		A		G	
T	TTT	Phe	TCt	Ser	TAT	Tyr	TGT	Cys
	TTC		TCc		TAC		TGC	
	TTA	Leu	TCa		TAA	Stop	TGA	Stop
	TTG		TCg		TAG		TGG	Trp
C	CTt	Leu	CCt	Pro	CAT	His	CGt	Arg
	CTc		CCc		CAC		CGc	
	CTa		CCa		CAA	Gln	CGa	
	CTg		CCg		CAG		CGg	
A	AT*t*	Ile	ACt	Thr	AAT	Asn	AGT	Ser
	AT*c*		ACc		AAC		AGC	
	AT*a*		ACa		AAA	Lys	AGA	Arg
	ATG	Met	ACg		AAG		AGG	
G	GTt	Val	GCt	Ala	GAT	Asp	GGt	Gly
	GTc		GCc		GAC		GGc	
	GTa		GCa		GAA	Glu	GGa	
	GTg		GCg		GAG		GGg	

FIG. 1.—Degeneracy of the genetic code. Nondegenerate sites are shown in uppercase, twofold-degenerate sites are in small caps, threefold-degenerate sites are in italics, and fourfold-degenerate sites are in lowercase.

replacements. The second positions of all 61 nontermination codons, as well as the first positions of 53 codons, belong to this class. The third positions of the methionine and tryptophan codons are also nondegenerate, and the third positions of the three isoleucine codons are actually threefold degenerate. Figure 1 displays the degeneracy classes for all 64 codons.

The existing methods proceed by first inferring the changes separating two homologous codons. All use a parsimony approach and choose pathways requiring the minimal number of nucleotide substitutions (although multiple-hit corrections are performed). With regard to estimating evolutionary trees, it has been shown that parsimony may be positively misleading in cases where evolutionary rates differ between lineages (Cavender 1978; Felsenstein 1978; Hendy and Penny 1989). The assignment of substitutions to degeneracy classes by considering only minimum-length pathways might be expected to suffer under the same conditions (large or unequal branch lengths) as does parsimony. If two codons differ by only a single site, there is a unique minimal pathway. When two or three sites differ, there are several minimal pathways, and each possibility must be weighted. Li et al. (1985*b*) suggest a method for weighting the pathways separating homologous codons that uses empirical data on the relative frequencies of codon substitutions. Miyata and Yasunaga (1980) suggest a weighting strategy based on physicochemical differences between amino acids. Perler et al. (1980) make no such corrections and treat all pathways as equally likely. It should be noted that all three methods will provide adequate and nearly equal estimates when substitution rates are low (Nei and Gojobori 1986). However, since none of the methods properly account for the possibility that a site's degeneracy class may have changed over time, and since none completely account for multiple substitutions at nucleotide sites, there is a need for an approach that fills these gaps. Such a method would be applicable to sequences of all levels of divergence.

Likelihood Estimation for Phylogenetic Trees

The new procedures suggested in this work rely heavily on the framework of Felsenstein (1981), who presented the first general and computationally effective method for estimating phylogenetic trees using the principle of maximum likelihood. His method will be illustrated with an example using four species. Figure 2 shows 1 of the 15 rooted trees for four species. Felsenstein assumed that nucleotide sites evolve independently, both of neighboring sites and of the homologous sites in other species, so that the overall likelihood is simply the product of individual site likelihoods. Let the four nucleotides A, C, G, and T be represented as 1, 2, 3 and 4, respectively, and let s_i be the nucleotide present in species i. The site will be indicated by context in order to simplify notation. Also, let π_k (k = 1, 2, 3, 4) denote the frequencies of the four nucleotides in a hypothetical "replacement pool." (The frequencies in this pool may differ because of properties of the region being analyzed. For instance, some genes have very high frequencies of G and C as compared with A and T. The π_k are usually taken to be the frequencies observed in the data.) The likelihood at site l of the tree in figure 2 can be written as

$$L_l(t) = \sum_{s_7} \sum_{s_6} \sum_{s_5} \{\pi_{s_7} P_{s_7 s_4}(t_4) P_{s_7 s_6}(t_6) \times P_{s_6 s_3}(t_3) P_{s_6 s_5}(t_5) P_{s_5 s_1}(t_1) P_{s_5 s_2}(t_2)\}, \quad (1)$$

and the total likelihood as

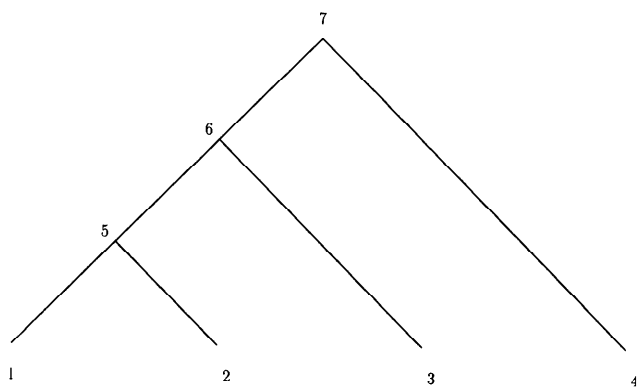

FIG. 2.—Rooted tree for four species

$$L_t = \prod_l L_l(t), \qquad (2)$$

where $P_{ij}(t)$ is the probability that a site initially with nucleotide i will be occupied by nucleotide j after a period of length t. $P_{ij}(t)$ will depend on the model of evolution chosen, as will the units of t. These details are important but are not needed at the moment, so discussion is postponed until later. Felsenstein (1981) provided several computational shortcuts, the most important (for our purposes) being the "pulley principle," which states that for $P_{ij}(t)$ arising from a time-reversible Markov process, there is no information as to the proper location of the root of the tree. In other words, likelihood calculations may be applied to unrooted trees, reducing the number of branches in the tree by one.

The computationally demanding part of this estimation method (for a known phylogeny) is maximizing the likelihood from equations (1) and (2) across all parameters. The number of parameters will depend on the substitution model chosen. For instance, there will be one parameter per branch for the model of Jukes and Cantor (1969) and two parameters per branch for the model of Kimura (1980). Any standard numerical optimization procedure can be used to find the maximum-likelihood estimates of these parameters.

Probabilistic Model for Codon Evolution

A likelihood approach for estimating nucleotide substitution rates is attractive, since it would avoid some of the problems mentioned previously. Development of a likelihood method requires that a model of sequence evolution be formulated. Note that the description of Felsenstein's likelihood-estimation procedure made no references to any particular model. Indeed, one of the procedure's strengths is that it can be used with many different evolutionary models. In previous work on relative-rate tests (Muse and Weir 1992), it was straightforward to use well-studied models of sequence evolution to form likelihood functions. This is not the case when substitution rates are divided into synonymous and nonsynonymous components. Simple models that treat nucleotide sites independently are inadequate because of codon structure, and this eliminates the bulk of published models. Methods have been suggested to model transition probabilities between amino acids (Adachi and Hasegawa 1992) and to provide empirical transition models for amino acids (Dayhoff et al. 1978), but these do not allow for estimation of synonymous substitution rates. Empirical data have also been collected to examine the relative frequencies of codon changes (Miyata and Yasunaga 1980; Li et al. 1985*b*), but inference may be limited to the evolutionary rates and degrees of divergence possessed by the genes and species found in the samples. Given that there is tremendous variation of substitution rates among genes (e.g., see Wu and Li 1985) and among lineages (e.g., see Gaut et al. 1992), it does not seem wise to use empirical data for these types of studies. It seems more appropriate to estimate the substitution rates for each set of data, rather than to assume that they are similar to rates from other studies.

Since no appropriate models for codon evolution exist in the literature, a new one must be formulated. The model is similar in form to those of Hasegawa et al. (1985) and Adachi and Hasegawa (1992) for the evolution of nucleotide and amino acid sequences. Define $P_{ij}(\alpha, \beta, dt)$ to be the probability of changing from codon i to codon j in a small amount of time dt. Numbering of the codons may be done in any convenient fashion. The parameters α and β determine the synonymous and nonsynonymous substitution rates, respectively. If we assume that in time dt only one nucleotide substitution event can occur in any particular codon, we can define the substitution process among the 61 nontermination codons as follows:

$$P_{ij}(\alpha, \beta, dt) = \begin{cases} \alpha\pi_n dt & \text{synonymous} \\ \beta\pi_n dt & \text{nonsynonymous} \\ 0 & \text{multiple substitutions needed} \end{cases},$$

where π_n is defined to be the equilibrium frequency of the "target nucleotide." (In practice, these values are taken to be the base frequencies observed in the data.) For example, the instantaneous probability of an AGG codon being substituted by an AGA codon would be $\alpha\pi_A dt$: the replacement of the G by an A results in no amino acid substitution, since both codons encode arginine. On the other hand, the instantaneous probability of an AGG codon being replaced by a CGA is zero. Even though the replacement is synonymous, it requires two nucleotide substitutions. Note that the model does not eliminate the possibility of such replacements; it simply requires that they occur through a series of steps, rather than in a single step. Termination codons are excluded from the model, since their occurrence would result in considerable terminal length variation in the functional proteins, a phenomenon that is easily detected when present.

A more realistic model would use the frequency of the target codon, or of the target amino acid, rather than that of the target nucleotide. The latter is incorporated here as a matter of convenience, since it greatly reduces the number of parameters that must be estimated. For most genes, there are currently insufficient data to de-

termine the frequencies of amino acids or codons with any degree of confidence. Each codon position seems to have its own distribution of amino acids, depending on its function. For instance, in membrane-spanning proteins it is of importance for some amino acids to be hydrophobic and for others to be hydrophilic. These two classes of sites would have very different codon distributions. Add to this the issue of codon-usage bias, the phenomenon of one codon being selectively favored over other codons encoding the same amino acid, and it is obvious that an accurate description of codon frequencies is very difficult.

Transition Probabilities

The maximum-likelihood method for estimating branch lengths in a phylogenetic tree requires calculation of transition probabilities for a given period of time, not just instantaneous probabilities. If we form a matrix, **A**, of instantaneous probabilities, $P_{ij}(\alpha, \beta, dt)$, the transition probabilities for time t are given by $e^{\mathbf{A}t}$ (Karlin and Taylor 1975). In previous studies of nucleotide sequences, explicit formulas for such probabilities were obtained; however, in this case the **A** matrix is 61×61, and there is a significant lack of symmetry (e.g., not all first and third positions are synonymous sites), so convenient expressions seem difficult, if not impossible, to find. The asymptotic frequencies, however, can be computed. Straightforward, but tedious, application of standard techniques shows that the asymptotic frequency of the codon consisting of nucleotides i, j, and k is $\pi_i\pi_j\pi_k/(1 - \Pi_{stop})$, where $\Pi_{stop} = \pi_T\pi_A\pi_G + \pi_T\pi_G\pi_A + \pi_T\pi_A\pi_A$. Recall that these quantities are needed for likelihood calculations. The transition probabilities for finite amounts of time are approximated by using the series expansion of $e^{\mathbf{A}t}$:

$$\mathbf{P}(t) = e^{\mathbf{A}t} = \mathbf{I} + \mathbf{A}t + (\mathbf{A}t)^2/2! + (\mathbf{A}t)^3/3! + \ldots . \quad (3)$$

Although it is certainly true that more sophisticated methods exist for evaluating $e^{\mathbf{A}t}$ (Moler and Van Loan 1978), the first 10 terms of this series seem to provide adequate approximations, providing nearly identical results in a shorter amount of time. Application of sparse matrix multiplication techniques (Press et al. 1992) tremendously decreases the computational burden of this expansion, since approximately 85% of the entries in **A** are zero. Once **P** has been approximated, we can evaluate the probability of the observed data given values of α, β, and t, and we can estimate parameters by using maximum likelihood. Only the products of substitution rates and time, αt and βt, are estimable.

Parameter Interpretation

A difficulty with this model, as well as with any model describing the evolution of coding sequences, is the interpretation of the parameters. This was alluded to by Muse and Weir (1992). The parameterizations prevalent in the literature deal with expected numbers of synonymous and nonsynonymous substitutions per nucleotide site. (Even in the simpler case of independent nucleotide substitution, the expected number of substitutions at a site depends on the unknown ancestral base, unless base frequencies are equal.) The parameterization used here does not easily allow such an interpretation. It can be argued that using expected numbers of substitutions per site is not all that meaningful. This is because of the lack of symmetry in the genetic code. The expected number of substitutions per site is not the same across sites, even when each site obeys the same evolutionary model. At a given point in time a site is either nondegenerate or two-, three-, or fourfold degenerate. Thus, some sites are more easily changed than others, and this is addressed by Li et al. (1985*b*). Even within a degeneracy class some positions are expected to undergo more substitutions over time than others, because of the nature of the genetic code. For instance, both the A in the ATG methionine codon and the A in the TAC tyrosine codon occupy nondegenerate sites. However, the first site in the methionine codon is only a single substitution away from becoming twofold degenerate. Substitution of the adenine base by either thymine or cytosine accomplishes this. The second position of the tyrosine codon will always remain nondegenerate. Even though the two nucleotides are currently in the same degeneracy class, one would expect more substitutions to occur at the first site of the methionine codon than at the second position of the tyrosine codon. Similar situations arise for fourfold-degenerate and nondegenerate sites. Another difficulty is that all nonsynonymous substitutions do not occur with equal probabilities. As an example, a change from Leu to Ile is much more probable than a change from Leu to Pro. This points out the fact that the "rate of nonsynonymous nucleotide substitution" is not a well-defined parameter.

Because the degeneracy class of a given site changes in a random fashion over time, measures such as the number of silent substitutions per silent site often may lose some of their meaning, particularly for long lineages when sites have been able to undergo one or more changes in degeneracy class. For this reason it seems best to consider the estimated parameters as simply the products of instantaneous substitution rates and time. This interpretation still lends itself easily to comparisons of substitution rates and of branch lengths, although it does not convey as strong a notion as to how many

events of each type (silent and replacement) have actually occurred or are expected in a lineage.

Standard mathematical approaches provide the form of the expected number of substitutions per site (averaged over all 61 codons), but the expression is complex and provides little intuitive value. To facilitate comparisons with other models and to demonstrate the difficulties mentioned above, it is desirable to give at least an approximation of the expected number of substitutions per site. Assume that the equilibrium probabilities of the four nucleotides are each ¼. If we choose a codon at random, of the nine possible single-nucleotide substitutions, an average of 2.197 are synonymous and 6.426 are nonsynonymous. To demonstrate this calculation, consider the ACT threonine codon. There are nine potential substitutions, three at each position. Of these, only the three third-site substitutions are synonymous; the remaining six are nonsynonymous. The numbers 2.197 and 6.426 were obtained by averaging the number of synonymous and nonsynonymous changes over all 61 codons. The numbers do not sum to nine because 23 substitution events lead to termination codons and are excluded from the count. That is, not all codons offer nine possible substitutions. The expected number of substitutions per codon after t units of time, $E_t(s)$ can be found as

$$E_t(s) = (¼) \int_0^t (2.197\alpha+6.426\beta)dt\,, \qquad (4)$$

so we can consider the expected number of substitutions at a particular codon to be approximately

$$(2.197\alpha t+6.426\beta t)/4\,. \qquad (5)$$

It is important to notice the number of assumptions and approximations required to arrive at this expression. The true expectation is a function of αt and βt, but, as mentioned above, the expression is complex. Evaluation of the expected number of substitutions per site is, however, straightforward when a computer is used. This suggests that the evolutionary model may be of some use in computing evolutionary distances.

Parameter Estimation

The difficult part of this exercise is the implementation. Since we must approximate the transition probabilities (which is quite slow because of the size of the matrices), and since we lack expressions for partial derivatives (although they could also be evaluated numerically), the maximum-likelihood estimation routine is doomed to be slow. Initially, a simulated annealing approach was implemented. This approach worked well but was very slow because of the large number of function evaluations needed, each of which requires computation of $e^{\mathbf{A}t}$ for either one or two values of **A**. Graphic analysis of the likelihood surface for several data sets suggested that no local maxima exist. In light of this evidence, parameters were maximized individually by bracketing the maximum and then using a modified bisection method with parabolic interpolation (Press et al. 1992). This process was applied repeatedly to each parameter in turn until the likelihood converged. The method seems to work quite well, finding the same estimates as the more robust simulated annealing method, in a fraction of the time.

Relative-Rate Tests

An important application for which this framework is useful is the comparison of substitution rates between lineages. Suppose that sequence data are available for three species, A, B, and O, where O is known from previous information to be an outgroup. The new model allows relative-rate tests to be performed on both synonymous and nonsynonymous rates in a straightforward and intuitive manner. By constraining the appropriate parameters in the maximization process, the following null hypotheses may be defined and tested (see fig. 3):

$$\text{LRS: } H_O\text{: } \alpha_A = \alpha_B$$
$$\text{LRN: } H_O\text{: } \beta_A = \beta_B$$
$$\text{LRB: } H_O\text{: } \alpha_A = \alpha_B,\ \beta_A = \beta_B\,.$$

The first test compares synonymous rates between the lineages leading to A and B, while the second test compares the nonsynonymous rates. The final test compares both synonymous and nonsynonymous rates simultaneously.

Previously (Wu and Li 1985; Muse and Weir 1992), these types of tests were phrased in terms of transition and transversion rates and then were applied to certain classes of sites (and assumptions invoked) in order to compare the rates of interest. Also, separate tests were performed on several classes of sites to investigate a single parameter. For instance, to compare synonymous rates, tests were performed on both fourfold- and twofold-degenerate sites. The present formulation avoids those contortions and actually makes more efficient use of the data by considering all possible pathways connecting two codons, each weighted by its appropriate probability. The relative-rate tests described by Muse and Weir (1992) are updated simply by using the substitution model described above. Also, rather than computing the likelihood by multiplying across nucleotide sites, the likelihood is now computed by multiplying

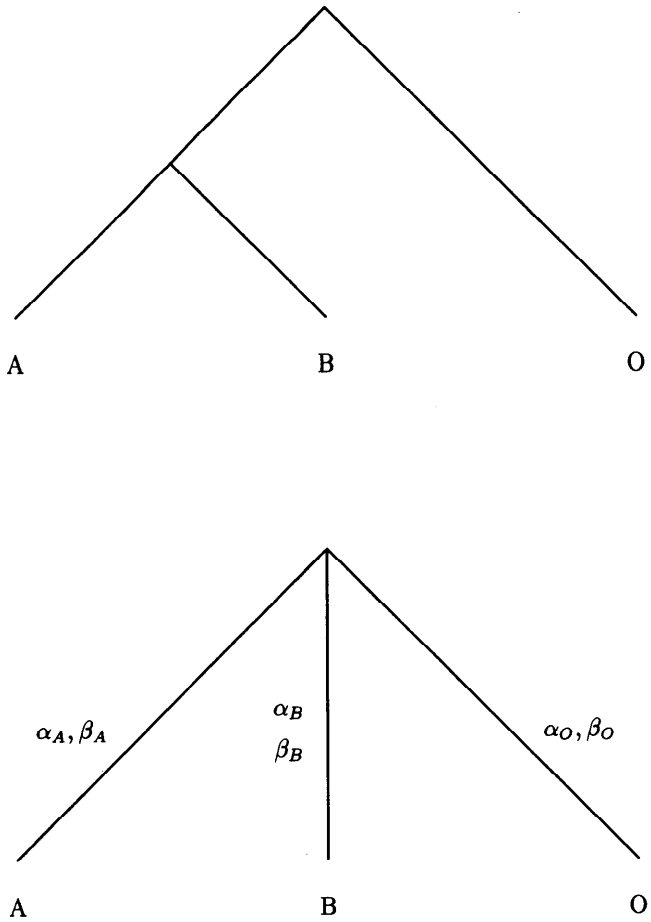

FIG. 3.—Rooted and unrooted trees for relative-rate test. α and β are synonymous and nonsynonymous substitution rates, respectively. The rates leading to sequences A and B are compared using an outgroup sequence, O.

across codons. A computer program implementing these three tests is available on request from the first author. The program is written in C for UNIX systems. Please send requests via electronic mail to SVM1@psuvm.psu.edu.

Simulation Results

In order to verify that the likelihood-ratio tests have the properties expected of them, two simulation experiments were performed. The size of the simulation study is small because of time constraints. In order to perform the three tests described above, approximately 35 min of CPU time are required on a Sun Sparc 2. Computation time is virtually independent of sequence length. The parameter values were chosen to be typical of loci used in the cpDNA work discussed in the next section.

The property most vital for these tests is that they follow the correct distribution under the null hypothesis of equal rates. A total of 250 replicate data sets were generated using $\alpha_A = \alpha_B = 1.0$, $\beta_A = \beta_B = 0.15$, $\alpha_O = 1.5$, and $\beta_O = 0.25$. The test statistic for the LRB test is predicted to have a χ^2_{2df} distribution. The remaining two tests should have a χ^2_{1df} distribution. The results in table 1 show that the observed frequencies of rejection are not significantly different from those predicted by the theoretical χ^2 distribution. The results also show that the relative-rate test formed by using the distance estimates of Nei and Gojobori (1986) behaves as expected under the null hypothesis.

For 100 additional replicates, the parameter values along branch B were altered: $\alpha_B = 0.75$ and $\beta_B = 0.10$. The tests should now reject the null hypothesis, since rates are accelerated along branch A. Not surprisingly, none of the tests had particularly high power, but the likelihood-ratio test for nonsynonymous rates appears to be more powerful than the corresponding test of Nei and Gojobori, at least for this set of parameter values. Note that neither test of synonymous rate heterogeneity seems to have any power for this degree of divergence. Both tests had rejection rates near the significance levels used. The issue of power, though it is certainly important, should not be overemphasized. The real advantages of the likelihood approach are the ability to avoid confounding factors such as degeneracy classes and its efficient use of data.

Although the present work is not immediately concerned with parameter estimation, the simulations do provide information on the maximum-likelihood estimators of αt and βt. Only estimates of the large synonymous rate parameter, α_O, showed any problems with bias. In both simulations the true value of α_O was 1.5. The average value of the maximum-likelihood estimate

Table 1
Simulation Results

	Equal Rates[b]			Unequal Rates[b]		
Test[a]	*n*	0.05	0.01	*n*	0.05	0.01
LRB	250	9	0	100	22	6
LRS	250	17	1	100	9	2
K_S	250	15	0	100	7	4
LRN	250	6	0	100	28	7
K_A	250	5	0	100	13	5

NOTE.—Two simulations were performed: 250 replicates where rates were equal in the two lineages, and 100 replicates where both synonymous and nonsynonymous rates were accelerated in one lineage.

[a] LRB, LRS, and LRN tests are the likelihood-ratio tests described in the text. K_S and K_A are tests of synonymous and nonsynonymous rate heterogeneity using the estimates of Nei and Gojobori (1986) in the relative-rate framework of Wu and Li (1985).

[b] Columns labeled "0.05" and "0.01" contain the number of replicate data sets for which the null hypothesis of equal rates was rejected at a 0.05 or 0.01 significance level.

was near 2.0 for the null case and near 1.75 for the alternative simulation. The likelihood surface is very flat in this region. The average values of the remaining estimates were very close to the true values.

Results

Relative Rates of Nucleotide Substitution in the Chloroplast Genome

The complete DNA sequence of the chloroplast genome has now been completed for three plant species: rice (*Oryza sativa*), tobacco (*Nicotiana tabacum*), and *Marchantia polymorpha.* These plants represent three very divergent lineages: the monocots, dicots, and liverworts, respectively. As might be expected, the phylogeny of these three species is well supported by fossil and morphological data. A phylogeny is shown in figure 4. Rice and tobacco diverged between 110 Mya (Dahlgreen et al. 1985) and 200 Mya (Wolfe et al. 1989). The liverworts and the vascular plants (including rice and tobacco) diverged 350–400 Mya (Wolfe et al. 1989). These data make it possible to perform relative-rate tests using loci from throughout the chloroplast genome. Rates in the rice and tobacco lineages can be compared using *Marchantia* as an outgroup.

Previous Results

The data used in this work were extracted from the complete chloroplast genome sequences present in the GenBank database (tobacco, Shinozaki et al. 1986; rice, Hiratsuka et al. 1989; and *Marchantia,* Ohyama et al. 1986). Gaut et al. (1993) have isolated and analyzed 38 coding sequences, using the likelihood-ratio test (LR2P) of Muse and Weir (1992). This test treats all nucleotide sites equally and tests the null hypothesis that both transition and transversion rates are equal between lineages. The third column of table 2 contains results of these tests using all three codon positions.

Since these species are very distantly related, silent sites seem to be saturated (results not shown). That is, so many substitution events have occurred in the evolutionary pathways connecting the species that no signal of common ancestry remains. For this reason, third-position sites were deleted from the analysis by Gaut et al. (1993). Column four of table 2 consists of tests performed on combined first and second positions. Three main observations arise from these tests: (i) rice loci have evolved at a faster rate than tobacco loci, suggesting lineage specific effects; (ii) highly accelerated loci are scattered throughout the chloroplast genome, suggesting that locus specific heterogeneity also exists; and (iii) there are clearly cases of heterogeneous nonsynonymous substitution rates.

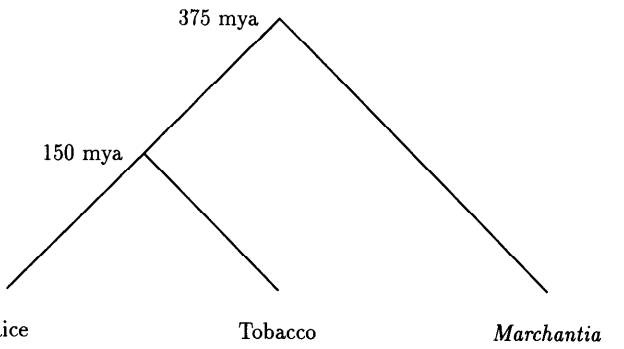

FIG. 4.—Phylogeny of rice, tobacco, and *Marchantia,* with approximate divergence times.

Tests for Differences in Synonymous and Nonsynonymous Substitution Rates

The work described above did not allow for an accurate characterization of rate heterogeneity, in the sense that it was difficult to designate rate differences as synonymous or nonsynonymous, because of the large divergence times. Up to 68% of codons are dissimilar in pairwise sequence comparisons, and among these variable codons up to 46% show differences at either two or three nucleotide sites. Although third positions were removed from the data in the previous work to reduce the number of silent sites, significant tests might still be attributed to differences in silent substitution rates. Some genes had significant results with the full data set but not with the reduced data set. It is tempting to attribute these to differences in synonymous rates, but this is not necessarily the case. To further investigate these issues, the new likelihood-ratio tests were employed. The results of these tests are found in columns five through seven of table 2.

Several observations can be made. First, little evidence for heterogeneity of synonymous substitution rates is found. Only three genes yield significant LRS tests, each at the 5% level. This may be because synonymous rates actually are equal, but it might also be attributed to saturation and the loss of statistical power that are due to the lengths of these lineages. Recall that the simulation results suggested that these tests have virtually no statistical power for this range of divergence. None of these three loci represent cases mentioned above, where the full nucleotide data led to rejection of the null but the reduced data did not. The tests do provide fairly widespread evidence of heterogeneous nonsynonymous substitution rates, both between lineages and among loci. Such results are not too surprising, given that Wolfe et al. (1987) found evidence of a tobacco slowdown and that both Bousquet et al. (1992) and Gaut et al. (1992) have described a grass acceleration. Rate heterogeneity between cpDNA regions has been documented by Wolfe

Table 2
Results of Relative-Rate Test

LOCUS (length in bp)	RELATIVE-RATE TEST[a]					
	LR2P(123)	LR2P(12)	LRB	LRS	LRN	R/T[b]
*atp*A (1,515)	≤0.01	≤0.01	≤0.01	0.05	≤0.01	R
*atp*B (1,497)	...	...	...	...	...	T
*atp*E (399)	≤0.01	≤0.01	≤0.01	...	≤0.01	R
*atp*F (540)	...	...	...	0.05	...	T
*atp*I (747)	≤0.01	...	0.05	0.05	...	R
*ndh*C (363)	...	...	...	...	...	R
*ndh*E (300)	0.05	≤0.01	≤0.01	...	≤0.01	R
*ndh*F (1,299)	...	...	...	...	...	T
*pet*A (963)	...	...	...	...	...	R
*pet*B (648)	...	...	...	...	0.05	T
*pet*D (483)	...	...	...	...	...	R
*psa*A (2,253)	...	...	...	...	0.05	R
*psa*B (2,205)	...	0.05	...	...	...	R
*psb*A (1,062)	...	...	...	...	0.05	R
*psb*B (1,527)	≤0.01	...	...	...	...	R
*psb*C (1,422)	...	...	...	...	...	T
*psb*D (1,062)	...	...	...	...	...	R
*psb*G (555)	...	...	...	...	...	R
*rbc*L (1,398)	...	...	...	...	...	R
*rpl*2 (825)	...	...	...	...	...	R
*rpl*14 (375)	...	...	...	...	...	R
*rpl*16 (432)	...	0.05	...	...	0.05	R
*rpl*20 (345)	...	0.05	...	...	...	R
*rpl*22 (252)	≤0.01	≤0.01	0.05	...	0.05	R
*rpl*23 (276)	...	...	...	...	...	R
*rpo*A (1,017)	...	...	...	...	...	R
*rpo*B (3,255)	≤0.01	≤0.01	≤0.01	...	≤0.01	R
*rpo*C1 (1,674)	≤0.01	≤0.01	≤0.01	...	≤0.01	R
*rpo*C2 (1,449)	0.05	...	...	...	...	R
*rps*2 (708)	≤0.01	≤0.01	≤0.01	...	≤0.01	R
*rps*3 (726)	...	0.05	0.05	...	0.05	R
*rps*4 (609)	...	0.05	...	...	0.05	R
*rps*7 (471)	...	...	0.05	...	≤0.01	R
*rps*8 (402)	...	...	...	...	...	R
*rps*11 (441)	≤0.01	≤0.01	≤0.01	...	≤0.01	R
*rps*12 (258)	0.05	...	...	...	...	R
*rps*14 (312)	...	...	...	...	...	T
*rps*19 (285)	...	...	0.05	...	≤0.01	R

[a] Entries are level at which the null hypothesis was rejected; an ellipsis (...) indicates that the null hypothesis was not rejected at the 0.05 level. LR2P(123) tests the null hypothesis of equal transition and transversion rates, ignoring codon structure. LR2P(12) is the same test but performed on data with third positions deleted. LRB tests the null hypothesis that both synonymous and nonsynonymous substitution rates are equal between rice and tobacco lineages. LRS and LRN test the null hypotheses that synonymous and nonsynonymous substitution rates (respectively) are equal between the two lineages.

[b] Letter indicate whether rice (R) or tobacco (T) was estimated to have evolved faster.

et al. (1987). Wolfe and Sharp (1988) documented saturation at silent sites when comparing liverwort and tobacco cpDNA sequences, as well as tremendous nonsynonymous rate variation between chloroplast loci.

The general pattern found here (considerable nonsynonymous rate heterogeneity but little or no difference in synonymous rates) is markedly different from the results reported by Gaut et al. (1992). In that study, substitution rates at the *rbc*L locus were compared for a large set of monocot species much more closely related than rice and tobacco. In fact, rice is more closely related to the outgroup used in that study (*Magnolia macrophylla*) than it is to tobacco. The findings suggested that synonymous rates were heterogeneous but provided little evidence for differences in nonsynonymous rates. It seems quite likely that the timescales are responsible for

the apparent inconsistencies. When closely related species are used, the number of silent substitutions remains low enough for relative-rate tests to retain the statistical power necessary to detect heterogeneity in synonymous rates. However, the number of replacement substitutions is excessively low, preventing tests from having power to detect differences in these rates. When very distantly related species are used, the situation is reversed. Power to detect differences in silent substitution rates is diminished because of saturation, but tests for heterogeneous nonsynonymous rates have enhanced power because of the accumulation of larger numbers of replacement events. Bousquet et al. (1992) also studied relative rates of substitution at the *rbc*L locus, but they used a more diverse set of taxa than did Gaut et al. (1992). These authors found evidence of both synonymous and nonsynonymous rate heterogeneity. This gives credence to the suggestion that the timescale is to some degree responsible both for the lack of significant differences in silent rates in this work and in that of Gaut et al. (1993) and for the lack of significant differences in nonsynonymous rates in the work of Gaut et al. (1992).

The tests confirm the findings of Gaut et al. (1993), who suggested that both lineage-specific and locus-specific effects on substitution rates existed. The final column in table 2 indicates whether the rice or tobacco lineage was estimated to have evolved faster. Of the 38 loci, 32 suggest that the rice gene evolved faster. If the two lineages had evolved at equal rates, one would expect half of the loci to favor each lineage. A sign test rejects the null hypothesis that the rice and tobacco lineages evolved with equal rates ($P<0.001$). For the loci yielding significant ($P<0.05$) LRB tests, the average ratio of rice branch lengths to tobacco branch lengths $((2.197\alpha t+6.426\beta t)/4)$ is 3.35. For the remaining loci the average is 1.8 (results not shown). These results suggest that not only does the entire rice chloroplast genome appear to have evolved faster than the tobacco chloroplast genome, but that there are also locus-specific effects on substitution rate as well. The latter is consistent with the findings of Wolfe and Sharp (1988).

Although a few results remain puzzling, it is comforting to see that the general patterns are quite consistent between the nucleotide and codon analyses. Both approaches appear to be capable of detecting rate heterogeneity when it is present. However, only the codon model allows one to determine accurately whether deviations are due to synonymous or nonsynonymous rates. An economical approach in practice might be to identify loci by using the tests from Muse and Weir (1992) or Wu and Li (1985) and then to use the tests from this work to describe the form of rate heterogeneity present.

Discussion

The problem of estimating and comparing synonymous and nonsynonymous substitution rates is a difficult one. The complex nature of the genetic code prevents simple models from being used in statistical procedures. Previous work in this area relied on the assumption that it is unlikely for multiple events to occur within the same codon during the time frame being studied. However, this makes application to distantly related taxa questionable. A likelihood approach based on a model of codon evolution avoids this difficulty. Problems with the stochastic nature of degeneracy classes are also handled properly using the likelihood framework. Although procedures based on models of codon evolution are computer intensive, they are applicable to any set of homologous coding regions, regardless of either the divergence times between taxa or the relative magnitudes of evolutionary rates, and provide satisfactory comparisons of substitution rates. The model presented in this work leads to intuitive procedures for performing relative-rate tests and also provides a framework for other procedures, such as estimating evolutionary distances and phylogenetic trees. The approach used in creating the model is also very flexible. Although we have not done so here, it seems possible to modify the A matrix to take into account physicochemical differences between amino acids. If one were to define a set of measures, d_{ij}, which indicated the similarity of codons i and j, the instantaneous probabilities could be altered by multiplying the off-diagonal elements by $d_{ij}/d_{\max}$. The distance measures of Grantham (1974) would be appropriate. Separate rates for transitions and transversions are also easily incorporated, but this may lead to difficulties in parameter estimation that are due to the absence of certain types of events in the data. Such modifications may be quite effective in compensating for any lack of realism present in the current model, particularly for use in phylogeny estimation problems.

Acknowledgments

This work has been supported in part by an NSF Graduate Fellowship to S.V.M., NIH postdoctoral fellowships GM16250 to S.V.M. and GM15528 to B.S.G., and by NIH grants GM32518 to North Carolina State University and GM45876 to Andrew G. Clark. Additional computing support was provided by the Pennsylvania State University Center for Computational Biology. Dr. Ron Gallant and Dr. John Monahan provided valuable advice on computational issues. Portions of this work were done in the course of SVM's Ph.D. studies. Dr. Bruce Weir deserves credit for the guidance provided during this work.

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1992. Amino acid substitution of proteins coded for in mitochondrial DNA during mammalian evolution. Jpn. J. Genet. **67**:187–197.

BOUSQUET, J., S. H. STRAUSS, A. H. DOERKESEN, and R. A. PRICE. 1992. Extensive variation in evolutionary rate of *rbc*L gene sequences among seed plants. Proc. Natl. Acad. Sci. USA **89**:7844–7848.

CAVENDER, J. A. 1978. Taxonomy with confidence. Math. Biosci. **40**:271–280. (erratum **44**:308).

DAHLGREEN, R. M. T., H. T. CLIFFORD, and P. F. YEO. 1985. The families of the Monocotyledons: structure, evolution, and taxonomy. Springer, Berlin.

DAYHOFF, M. O., R. M. SCHWARTZ, and B. C. ORCUTT. 1978. A model of evolutionary change in proteins. Pp. 345–352 *in* M. O. Dayhoff, ed. Atlas of protein sequence and structure. Vol. **5**, Suppl. 3. National Biomedical Research Foundation, Washington, D.C.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**: 401–410.

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

GAUT, B. S., S. V. MUSE, W. D. CLARK, and M. T. CLEGG. 1992. Relative rates of nucleotide substitution at the *rbc*L locus of monocotyledonous plants. J. Mol. Evol. **35**:292–303.

GAUT, B. S., S. V. MUSE, and M. T. CLEGG. 1993. Relative rates of nucleotide substitution in the chloroplast genome. Mol. Phylogenet. Evol. **2**:89–96.

GRANTHAM, R. 1974. Amino acid difference formula to help explain protein evolution. Science **185**:862–864.

HASEGAWA, M., H. KISHINO, and T. YANO. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

HENDY, M. D., and D. PENNY. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. **38**:297–309.

HIRATSUKA, J., H. SHIMADA, R. WHITTIER, T. ISHIBASHI, M. SAKAMOTO, M. MORI, C. KONDO, Y. HONJI, C.-R. SUN, B.-Y. MING, Y.-Q. LI, A. KANNO, Y. NISHIZAWA, A. HIRAI, K. SHINOZAKI, and M. SUGIURA. 1989. The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid inversion during the evolution of the cereals. Mol. Gen. Genet. **217**:185–194.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. M. Munro, ed. Mammalian protein metabolism. Academic Press, New York.

KARLIN, S., and H. M. TAYLOR. 1975. A first course in stochastic processes, 2d ed. Academic Press, New York.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

LI, W.-H., C.-C. LUO, and C.-I. WU. 1985*a*. Evolution of DNA sequences. Pp. 1–94 *in* R. J. MacIntyre, ed. Molecular evolutionary genetics. Plenum, New York.

LI, W.-H., C.-I. WU, and C.-C. LUO. 1985*b*. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**:150–174.

MIYATA, T., and T. YASUNAGA. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitution from homologous sequences and its application. J. Mol. Evol. **16**:23–26.

MOLER, C., and C. VAN LOAN. 1978. Nineteen dubious ways to compute the exponential of a matrix. SIAM Rev. **20**: 801–836.

MUSE, S. V., and B. S. WEIR. 1992. Testing for equality of evolutionary rates. Genetics **132**:269–276.

NEI, M., and T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

OHYAMA, K., H. FUKUZAWA, T. KOHCHI, H. SHIRAI, T. SANO, S. SANO, K. UMESONO, Y. SHIKI, M. TAKEUCHI, Z. CHANG, S. AOTA, H. INOKUCHI, and H. OZEKI. 1986. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature **322**:572–574.

PERLER, R., A. EFSTRATIADIS, P. LOMEDICO, W. GILBERT, R. KLODNER, and J. DODGSON. 1980. The evolution of genes: the chicken preproinsulin gene. Cell **20**:555–566.

PRESS, W. H., S. A. TEUKOLSKY, W. T. VETTERLING, and B. P. FLANNERY. 1992. Numerical recipes, 2d ed. Cambridge University Press, Cambridge.

SHINOZAKI, K., M. OHME, M. TANAKA, T. WAKASUGI, N. HAYASHIDA, T. MATSUBAYASHI, N. ZAITA, J. CHUMWONGSE, J. OBOKATA, K. YAMAGUCHI-SHINOZAKI, C. OHTO, K. TORAZAWA, B.-Y. MENG, M. SUGITA, H. DENO, T. KAMOGASHIRA, K. YAMADA, J. KUSUDA, F. TAKAIWA, A. KATO, N. TOHDOD, H. SHIMADA, and M. SUGIURA. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J. **5**:2043–2049.

WOLFE, K. H., M. GOUY, Y.-W. YANG, P. M. SHARP, and W.-H. LI. 1989. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. Proc. Natl. Acad. Sci. USA **86**:6201–6205.

WOLFE, K. H., W.-H. LI, and P. M. SHARP. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. USA **84**:9054–9058.

WOLFE, K. H., and P. M. SHARP. 1988. Identification of functional open reading frames in chloroplast genomes. Gene **66**:215–222.

WU, C.-I., and W.-H. LI. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA **82**:1741–1745.

PAUL SHARP, reviewing editor

Received September 29, 1993

Accepted April 18, 1994