Likelihood Ratio Tests for Detecting Positive Selection and Application to Primate Lysozyme Evolution

Ziheng Yang

Department of Biology, University College, London

An excess of nonsynonymous substitutions over synonymous ones is an important indicator of positive selection at the molecular level. A lineage that underwent Darwinian selection may have a nonsynonymous/synonymous rate ratio (d_N/d_S) that is different from those of other lineages or greater than one. In this paper, several codon-based likelihood models that allow for variable d_N/d_S ratios among lineages were developed. They were then used to construct likelihood ratio tests to examine whether the d_N/d_S ratio is variable among evolutionary lineages, whether the ratio for a few lineages of interest is different from the background ratio for other lineages in the phylogeny, and whether the d_N/d_S ratio for the lineages of interest is greater than one. The tests were applied to the lysozyme genes of 24 primate species. The d_N/d_S ratios were found to differ significantly among lineages, indicating that the evolution of primate lysozymes is episodic, which is incompatible with the neutral theory. Maximum-likelihood estimates of parameters suggested that about nine nonsynonymous and zero synonymous nucleotide substitutions occurred in the lineage leading to hominoids, and the d_N/d_S ratio for that lineage is significantly greater than one. The corresponding estimates for the lineage ancestral to colobine monkeys were nine and one, and the d_N/d_S ratio for the lineage is not significantly greater than one, although it is significantly higher than the background ratio. The likelihood analysis thus confirmed most, but not all, conclusions Messier and Stewart reached using reconstructed ancestral sequences to estimate synonymous and nonsynonymous rates for different lineages.

Introduction

As mutation and selection have different effects on synonymous (d_S) and nonsynonymous (d_N) substitution rates, estimation of these rates provides an important means for understanding the mechanisms of molecular sequence evolution. A d_N/d_S ratio significantly greater than one is a convincing indicator of positive selection. Using this idea, Messier and Stewart (1997) performed an interesting analysis of adaptive evolution in primate lysozymes. They used parsimony and likelihood methods to reconstruct lysozyme genes of extinct ancestors in a phylogeny of primates. The reconstructed and observed sequences were then used to estimate the numbers of synonymous and nonsynonymous substitutions per site (d_S and d_N) along each branch in the phylogenetic tree, with the approximate method of Li (1993) used for pairwise sequence comparison. Their analysis identified two lineages with elevated d_N/d_S ratios, indicating episodes of positive Darwinian selection in the lysozyme. One lineage, expected from previous analysis (Stewart, Schilling, and Wilson 1987), is ancestral to colobine monkeys, which have foreguts in which lysozyme is present and has acquired a new digestive function. Another lineage that is ancestral to the hominoids was previously unsuspected.

Although ancestral character states (nucleotides or amino acids, for example) reconstructed using parsimony have been used in all sorts of analyses as if they were observed data (see Maddison and Maddison 1992 for a review), this is not a rigorous statistical approach. Since the reconstructed sequences involve random errors and systematic biases (e.g., Collins, Wimberger, and Naylor 1994; Perna and Kocher 1995), inferences based on such pseudodata may be unsafe. The sequences analyzed by Messier and Stewart (1997) are very similar, so the accuracy of ancestral reconstruction is expected to be high (Messier and Stewart 1997; Yang, Kumar, and Nei 1995). Nevertheless, the authors' results should be examined by a more rigorous statistical analysis.

It is possible, and also advantageous, to avoid using reconstructed ancestral sequences to estimate d_S and d_N for lineages in the phylogeny. This can be achieved by adopting a likelihood approach, which averages over all possible ancestral sequences at each interior node in the tree, weighted appropriately according to their relative likelihoods of occurrence. Furthermore, in a likelihood model, it is straightforward to take into account the transition/transversion rate bias and nonuniform codon usage, factors that are not properly accommodated in approximate methods of pairwise comparison but are nevertheless very important in the estimation of d_S and d_N (Ina 1995). Thus, the likelihood method has the important advantage of being based on more realistic models of sequence evolution. In this paper, I develop several codon-based likelihood models that allow for different d_N/d_S ratios among evolutionary lineages. The models are then used to construct likelihood ratio tests to examine various hypotheses. The tests are applied to the lysozyme data of Messier and Stewart (1997).

Data and Methods

Data

The lysozyme gene sequences of 24 primate species analyzed by Messier and Stewart (1997) are used. The phylogenetic tree of the species is shown in figure 1, and used in later analysis. Some sequences are identical, and only the $n = 19$ distinct sequences are used in the analysis of this paper. Use of all 24 sequences

Key words: positive selection, episodic evolution, neutral theory, lysozymes, primates, maximum likelihood, likelihood ratio test, synonymous rates, nonsynonymous rates.

Address for correspondence and reprints: Ziheng Yang, Department of Biology (Galton Lab), 4 Stephenson Way, London NW1 2HE, United Kingdom. Email: z.yang@ucl.ac.uk.

Mol. Biol. Evol. 15(5):568–573. 1998

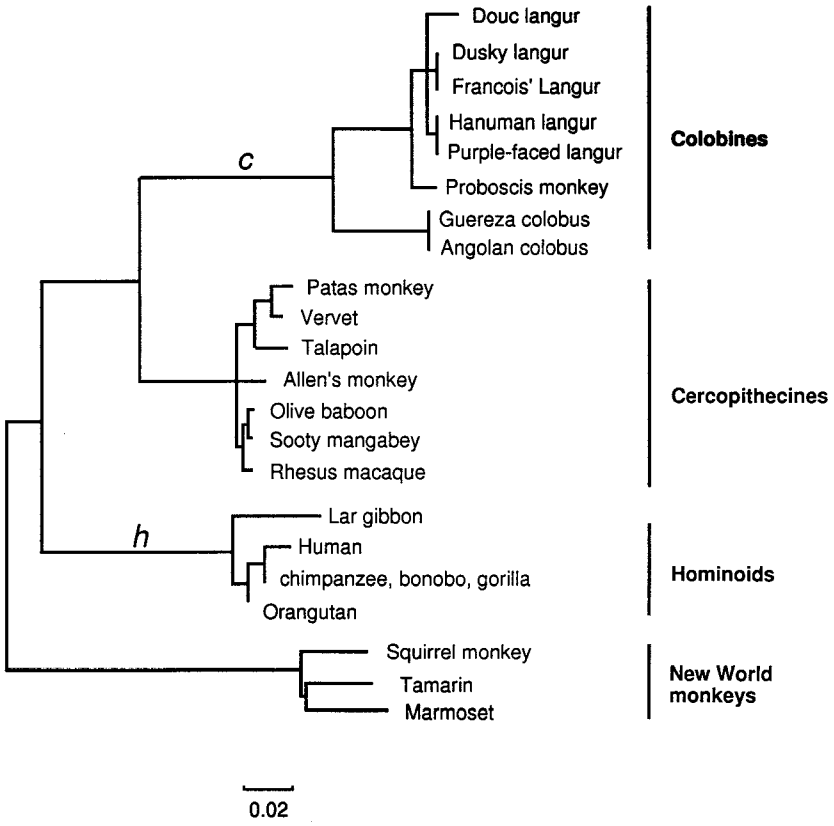

FIG. 1.—Phylogeny of the primate species analyzed by Messier and Stewart (1997) and this paper. Some species have identical lysozyme gene sequences, and only the 19 distinct sequences are used. Branches are drawn in proportion to their lengths, defined as the expected numbers of nucleotide substitutions per codon and estimated using the one-ratio model, which assumes the same d_N/d_S ratio for all branches in the tree (Goldman and Yang 1994). The tree topology, but not the branch lengths, is used to fit different models.

would introduce several zero-length branches and would not change any of the results to be presented below. The genes are quite similar among species; out of the 103 codon sites, 82 are constant. (A codon site is said to be constant only if all the three nucleotides are identical across all species.) A subset of $n = 7$ species, identified in figure 2 and selected to represent the four major groups in the phylogeny, is also analyzed. This subset is referred to as the "small" data set, while the complete data set is referred to as the "large" data set. Differences between the analyses of the two data sets will give us an indication of the sensitivity of the results to species

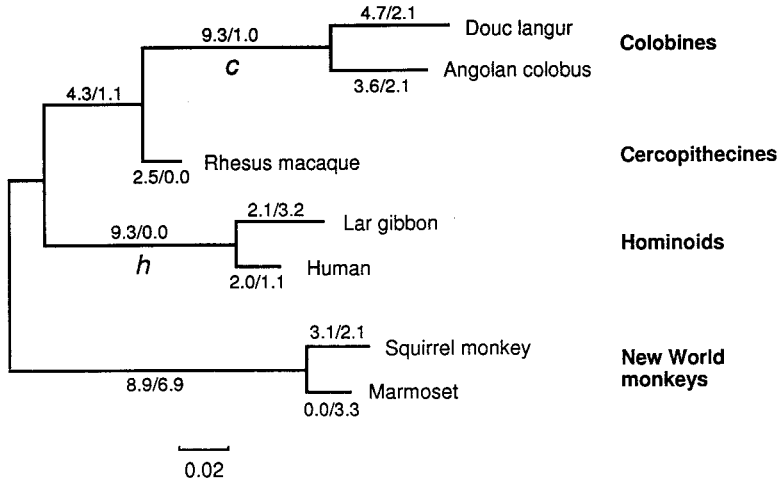

FIG. 2.—Phylogeny of a subset of seven primate species selected from those of figure 1 to represent the four major groups of species. The two numbers shown along each branch are the maximum-likelihood estimates of the numbers of nonsynonymous and synonymous substitutions for the entire gene along that branch. The "free-ratio" model is used, which assumes a different d_N/d_S ratio (parameter ω) for each branch in the tree. Branches are drawn in proportion to estimates of their lengths.

sampling. Branches *c* (ancestral to colobines) and *h* (ancestral to hominoids) in figures 1 and 2 were suggested by Messier and Stewart (1997) to be potentially under positive selection and will be the lineages of interest in later analysis.

Methods

The basic model for the likelihood analysis is a version of the codon-substitution model of Goldman and Yang (1994) and accounts for the genetic code structure, transition/transversion rate bias, and different base frequencies at codon positions. The instantaneous substitution rate from codon i to codon j ($i \neq j$) is specified as

$$q_{ij} = \begin{cases} 0 & \text{if the two codons differ at more than one position,} \\ \pi_j & \text{for synonymous transversion,} \\ \kappa\pi_j & \text{for synonymous transition,} \\ \omega\pi_j & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j & \text{for nonsynonymous transition,} \end{cases} \quad (1)$$

where κ is the transition/transversion rate ratio, ω is the nonsynonymous/synonymous rate ratio, and π_j is the equilibrium frequency of codon *j*, calculated from the nucleotide frequencies at the three codon positions. Note that under this model, $\omega = d_N/d_S$.

Several models can be constructed that allow for different levels of heterogeneity in the d_N/d_S ratio among lineages. The simplest model assumes the same ratio for all branches in the phylogeny, and will be referred to as the ''one-ratio'' model. The most general model assumes an independent d_N/d_S ratio for each branch in the phylogeny. This model, referred to as the ''free-ratio'' model, involves as many ω parameters as the number of branches in the tree and is parameter-rich for a tree of many species. It will be applied to the small data set only. Many intermediate models that lie between the two extremes are possible and are implemented in the PAML program package (Yang 1997; see below for program performance and availability). Models used in this paper, for the analysis of the lysozyme genes of figures 1 and 2, make different assumptions about the d_N/d_S ratios for branches *h* and *c*, relative to the background d_N/d_S ratio (ω_0) for all other branches. For instance, the ''two-ratio'' model assumes that the branches of interest (branch *h* or *c* or both) have a d_N/d_S ratio (ω_1) that is different from the background ratio ω_0. The ''three-ratio'' model assumes that the ratios for branches *h* and *c* are different and both are different from the background ratio.

The above models can be compared using the likelihood ratio test to examine interesting hypotheses. For example, the likelihood values under the one-ratio and free-ratio models can be compared to test whether the d_N/d_S ratios are different among lineages. Heterogeneity in the d_N/d_S ratio among lineages may be caused by positive selection or relaxed selectional constraints in some lineages; in either case, the neutral model of sequence evolution is violated. The one-ratio and two-ratio models can be compared to examine whether the lineages of interest have a different d_N/d_S ratio from other lineages. The likelihood values under the two-ratio model with and without the constraint $\omega_1 \leq 1$ can be compared to test whether the branches of interest have a d_N/d_S ratio that is greater than one (i.e., whether $\omega_1 > 1$). This test directly examines the possibility of positive selection operating on specific lineages.

The reader is referred to Kendall and Stuart (1979; Chapters 18 and 24) for an accessible account of the mathematical theory of maximum-likelihood estimation and the likelihood ratio test. Calculation of the likelihood function under the variable-ratio models in this paper can be adapted from the algorithm for the one-ratio model developed previously (Goldman and Yang 1994; see also Felsenstein 1981). A standard numerical algorithm is used to obtain the eigenvalues and eigenvectors of the rate matrix, $Q = \{q_{ij}\}$, to be used for calculating the transition probability matrix for a branch of length *t*, $P(t) = e^{Qt}$. When the d_N/d_S ratio (parameter ω) is different from branch to branch, it is necessary to perform this calculation for each branch. Parameters in the model including branch lengths of the tree, the transition/transversion rate ratio (κ), and the d_N/d_S ratio(s) (ω parameters) are estimated by maximum likelihood. A numerical iteration algorithm is used for this purpose. Different starting values are used in the iteration to guard against the existence of multiple local optima, and the likelihood function appeared to be well-behaved and unimodal. This means that the estimate of ω_1 for the branches of interest under the constraint (null hypothesis) $\omega_1 \leq 1$ is obtained by fixing $\omega_1 = 1$ if the estimate without the constraint is >1. The codon frequency parameters (π_j's in equation 1) are calculated using the observed nucleotide frequencies at the three codon positions.

Results

Test of Variable d_N/d_S Ratios Among Lineages

The free-ratio model, which assumes a different ω parameter for each branch in the tree, is applied to the small data set only (fig. 2). The log likelihood value under this model is $l_1 = -896.41$. The one-ratio model, which assumes the same ω parameter for the entire tree, leads to $l_0 = -906.02$. Since the free-ratio model involves 11 ω parameters for the 11 branches, while the one-ratio model assumes one, twice the log likelihood difference, $2\Delta l = 2(l_1 - l_0) = 19.21$, can be compared with a χ^2 distribution with df = 10 to test whether the free-ratio model fits the data significantly better than the one-ratio model. The difference between the two models is significant ($0.01 < P < 0.05$), indicating that the d_N/d_S ratios are indeed different among lineages. This conclusion also holds when the sequence of patas monkey (another cercopithecine monkey) is added to the analysis, and when codon frequencies (π_j in eq. 1) are used as free parameters in the model (Goldman and Yang 1994).

Maximum-likelihood estimates of parameters for each branch under the free-ratio model (*t* for branch length and ω for the d_N/d_S ratio), together with the estimated κ (4.59), can be used to calculate the numbers of synonymous and nonsynonymous substitutions per

Table 1
Log Likelihood Values and Parameter Estimates Under Different Models

Model	p	ℓ	$\hat{\kappa}$	$\hat{\omega}_0$	$\hat{\omega}_H$	$\hat{\omega}_C$
Large data set (n = 19)						
A. One ratio: $\omega_0 = \omega_H = \omega_C$	35	−1043.84	4.157	0.574	$=\hat{\omega}_0$	$=\hat{\omega}_0$
B. Two ratios: $\omega_0 = \omega_H, \omega_C$	36	−1041.70	4.163	0.489	$=\hat{\omega}_0$	3.383
C. Two ratios: $\omega_0 = \omega_C, \omega_H$	36	−1039.92	4.186	0.484	∞	$=\hat{\omega}_0$
D. Two ratios: $\omega_0, \omega_H = \omega_C$	36	−1037.59	4.199	0.392	7.166	$=\hat{\omega}_H$
E. Three ratios: $\omega_0, \omega_H, \omega_C$	37	−1037.04	4.196	0.392	∞	3.516
F. Two ratios: $\omega_0 = \omega_H, \omega_C = 1$	35	−1042.50	4.074	0.488	$=\hat{\omega}_0$	1
G. Two ratios: $\omega_0 = \omega_C, \omega_H = 1$	35	−1042.29	4.058	0.484	1	$=\hat{\omega}_0$
H. Two ratios: $\omega_0, \omega_H = \omega_C = 1$	35	−1040.32	3.974	0.392	1	1
I. Three ratios: $\omega_0, \omega_H, \omega_C = 1$	36	−1037.92	4.101	0.392	∞	1
J. Three ratios: $\omega_0, \omega_H = 1, \omega_C$	36	−1039.49	4.063	0.392	1	3.448
Small data set (n = 7)						
A. One ratio: $\omega_0 = \omega_H = \omega_C$	13	−906.02	4.540	0.807	$=\hat{\omega}_0$	$=\hat{\omega}_0$
B. Two ratios: $\omega_0 = \omega_H, \omega_C$	14	−904.64	4.561	0.686	$=\hat{\omega}_0$	3.506
C. Two ratios: $\omega_0 = \omega_C, \omega_H$	14	−903.08	4.568	0.675	∞	$=\hat{\omega}_0$
D. Two ratios: $\omega_0, \omega_H = \omega_C$	14	−901.63	4.605	0.540	7.263	$=\hat{\omega}_H$
E. Three ratios: $\omega_0, \omega_H, \omega_C$	15	−901.10	4.598	0.540	∞	3.646
F. Two ratios: $\omega_0 = \omega_H, \omega_C = 1$	13	−905.48	4.437	0.686	$=\hat{\omega}_0$	1
G. Two ratios: $\omega_0 = \omega_C, \omega_H = 1$	13	−905.38	4.413	0.675	1	$=\hat{\omega}_0$
H. Two ratios: $\omega_0, \omega_H = \omega_C = 1$	13	−904.36	4.312	0.543	1	1
I. Three ratios: $\omega_0, \omega_H, \omega_C = 1$	14	−902.02	4.465	0.541	∞	1
J. Three ratios: $\omega_0, \omega_H = 1, \omega_C$	14	−903.48	4.435	0.541	1	3.559

NOTE.—p, number of parameters in the model not including the nine parameters for codon frequencies (π_j's in eq. 1). Parameters ω_H, ω_C, and ω_0 are the d_N/d_S ratios for branches h, c, and all other branches, respectively (see figs. 1 and 2). Estimates of branch lengths are not shown.

site (d_S and d_N) for that branch (Goldman and Yang 1994). According to the model of equation 1 and the approach of Goldman and Yang (1994), the lysozyme gene has 107.9 synonymous and 282.1 nonsynonymous sites (with a total of 390 sites or 130 codons). The numbers of synonymous and nonsynonymous substitutions for the entire gene along each branch can then be calculated as the product of the rate per site (d_S or d_N) and the number of sites in the gene. For example, the estimates for branch c (fig. 2) are $\hat{t} = 0.079$ (nucleotide substitutions per codon) and $\hat{\omega} = 3.512$, which lead to $d_N = 0.033$ and $d_S = 0.009$. Thus, 9.3 (=0.033 × 282.1) nonsynonymous and 1.0 (=0.009 × 107.9) synonymous substitutions are estimated to have occurred along branch c. The estimates for branch h are $\hat{t} = 0.071$ and $\hat{\omega} = \infty$, yielding $d_N = 0.033$ and $d_S = 0$. These estimates suggest that 9.3 nonsynonymous and 0 synonymous substitutions occurred along branch h, with an even more extreme d_N/d_S ratio than for branch c. Estimates for other branches are shown in figure 2. Branches c and h differ from other branches in the tree in that both of them are long (i.e., have accumulated many changes) and have very high d_N/d_S ratios. These results are similar to the findings of Messier and Stewart (1997), although estimates of the d_N/d_S ratios are somewhat different in the two analyses.

While estimates of the d_N/d_S ratios for branches other than h and c are not identical, most of these background branches are very short (with very few changes) and do not contain much information. Changing the d_N/d_S ratios for these branches will cause little change in the likelihood under the model. It therefore appears justifiable to assume the same d_N/d_S ratio (ω_0) for those background branches, as we will do in later analyses in this paper.

Test of Positive Selection Along the Ancestral Lineages of Hominoids and Colobines

Log likelihood values and maximum-likelihood estimates of parameters under different models are given in table 1. Both the large and small data sets are analyzed. The models place different constraints on the three d_N/d_S ratio parameters: ω_H for branch h, ω_C for branch c, and ω_0 for all other (background) branches (see figs. 1 and 2). The simplest model assumes one d_N/d_S ratio (table 1, A and F), while the most general model in table 1 (E, I, and J) assumes three ratios. The three possible two-ratio models (B–D and F–H) are also fitted to the lysozyme data. For example, model B assumes that branch c has a different ratio (ω_C) while all other branches have the same background ratio ($\omega_H = \omega_0$). In models F–J, the d_N/d_S ratio for the branch(es) of interest is fixed at one.

We examine the maximum-likelihood estimates of parameters first. Estimates of the transition/transversion rate ratio are quite similar under different models, and range from 4.0 to 4.2 for the large data set and from 4.3 to 4.6 for the small data set. The estimate of the d_N/d_S ratio under the one-ratio model ($\omega_0 = \omega_H = \omega_C$) is 0.57 for the large data set and 0.81 for the small data set. The difference is due to the fact that the lineages not included in the tree for the small data set (fig. 2) involve low d_N/d_S ratios. Both estimates are smaller than one and indicate that, on average, synonymous substitutions occur more often than nonsynonymous substitutions and that lysozyme has spent a majority of time under neg-

Table 2
Likelihood Ratio Statistics ($2\Delta\ell$) for Testing Hypotheses

Null Hypothesis Tested	Assumption Made	Models Compared	Large Data Set ($n = 19$)	Small Data Set ($n = 7$)
A. $(\omega_H = \omega_C) = \omega_0$	$\omega_H = \omega_C$	A and D	12.50**	8.78**
B. $\omega_C = \omega_0$	$\omega_H = \omega_0$	A and B	4.28*	2.76
C. $\omega_C = \omega_0$	ω_H free	C and E	5.76*	3.96*
D. $\omega_H = \omega_0$	$\omega_C = \omega_0$	A and C	7.84**	5.88*
E. $\omega_H = \omega_0$	ω_C free	B and E	9.32**	7.08**
A′. $(\omega_H = \omega_C) \leq 1$ ..	$\omega_H = \omega_C$	D and H	5.46*	5.46*
B′. $\omega_C \leq 1$	$\omega_H = \omega_0$	B and F	1.60	1.68
C′. $\omega_C \leq 1$	ω_H free	E and I	1.76	1.84
D′. $\omega_H \leq 1$	$\omega_C = \omega_0$	C and G	4.74*	4.60*
E′. $\omega_H \leq 1$	ω_C free	E and J	4.90*	4.76*

* Significant ($P < 5\%$; $\chi^2_1 = 3.84$).
** Extremely significant ($P < 1\%$; $\chi^2_1 = 6.63$).

ative selection during primate evolution. Estimates of ω_C for branch *c* range from 3.4 to 3.6 when ω_C is free to vary (models B, E, and J in table 1). Estimates of ω_H are always infinite when ω_H is assumed to be a free parameter (models C, E, and I), indicating the absence of synonymous substitutions along branch *h*. Estimates of the d_N/d_S ratios for the two branches are thus highly similar across models and data sets and are also almost identical to estimates obtained under the free-ratio model for the small data set (3.5 for ω_C and ∞ for ω_H). When the same ratio is assumed for branches *h* and *c* ($\omega_H = \omega_C$; model D), the estimate is 7.2, which is an average over the two branches. Estimates of the background d_N/d_S ratio (ω_0) are 0.39 and 0.54 for the large and small data sets, respectively, when ω_H and ω_C are not constrained to be equal to ω_0 (models D, E, I, and J).

The log likelihood values under different models of table 1 were compared to test various hypotheses, with the results shown in table 2. Tests A–E examine whether the d_N/d_S ratio for the branch(es) of interest is different from (that is, greater than) the background ratio, while tests A′–E′ examine whether the ratio is greater than one. In tests A and A′, branches *h* and *c* are those of interest and are assumed to have the same d_N/d_S ratio. This ratio ($\omega_H = \omega_C$) is found to be significantly greater than the background ratio ω_0 ($P < 1\%$; table 2, A) and also significantly greater than one ($P < 5\%$; table 2, A′). To find out which of branches *h* and *c* may have caused the significant results, ω_H and ω_C are allowed to differ in tests B–E and B′–E′; in each case, only one of the two is compared with the background ratio ω_0, while the other is either allowed to vary freely or constrained to be equal to ω_0. These tests suggest that ω_H is significantly greater than the background ratio ω_0 ($P < 1\%$; table 2, D and E) and also significantly greater than one ($P < 5\%$; table 2, D′ and E′). Similar tests suggest that ω_C is significantly greater than ω_0 ($P < 5\%$; table 2, B and C for the large data set and C for the small data set) except for the small data set under the constraint $\omega_H = \omega_0$, for which $P = 10\%$ (table 2, B for the small data set). However, ω_C is not significantly greater than one whether or not ω_H is constrained to be equal to ω_0 (*P* ranges from 17% to 20%; table 2, B′ and C′).

It is remarkable that the large and small data sets lead to the same conclusions in 19 out of 20 comparisons (table 2). The test statistics ($2\Delta l$) for tests A′–E′ are highly similar between the two data sets. The statistics for tests A–E are more different, obviously because the inclusion of additional lineages with low d_N/d_S ratios in the large tree (fig. 1) makes it more obvious that ω_H and/or ω_C are greater than the background ratio ω_0. It is also noted that the results of the tests are not sensitive to the common assumption made in the null and alternative hypotheses about the d_N/d_S ratios. For instance, ω_H is significantly greater than one whether or not the constraint $\omega_C = \omega_0$ is enforced, and ω_C is not significantly greater than one whether or not $\omega_H = \omega_0$ is assumed. The results are thus robust to species sampling and to minor changes of assumptions made in the model.

Discussion

Maximum-likelihood analyses in this paper showed that the d_N/d_S ratios in the primate lysozyme genes are highly variable among evolutionary lineages, indicating that the evolution of primate lysozymes is incompatible with a neutral model of sequence evolution. The d_N/d_S ratio along branch *h* (the lineage leading to hominoids) was found to be significantly greater than the background ratio for other lineages and also significantly greater than one, indicating that positive selection may have operated during the lysozyme evolution along this lineage. These results are in agreement with those of Messier and Stewart (1997). The likelihood ratio tests also suggested that the d_N/d_S ratio for branch *c* (the lineage leading to colobine monkeys) was significantly greater than the background ratio but not significantly greater than one. This result is somewhat different from that of Messier and Stewart (1997), who found that both ω_H and ω_C were significantly greater than one. The two analyses are different in many respects (see *Introduction*), and it is not very clear which factor is the major cause of this discrepancy.

Before we draw conclusions about possible positive selection or lack of it in the evolution of primate lysozymes, particularly along branch *c,* several factors should be considered. First, strictly speaking, a proper statistical test requires the null hypothesis to be specified before the data are analyzed. This is not entirely the case in the analysis in this paper, since branch *h* was identified in Messier and Stewart's (1997) analysis of the same data set and used together with branch *c* as the branches of interest in the likelihood ratio tests of this paper. While the lack of independence of the hypothesis on data is not expected to have a great effect, it does increase the probability of rejecting the null hypothesis. Furthermore, multiple tests were performed in table 2, yet it is well known that multiple tests using the same data may lead to significant results simply because of chance effect. This problem, however, does not appear to be serious in this paper. The primary question has been whether ω_H and ω_C are greater than one, and most

tests in table 2 serve as robustness analyses only. It might be argued that not many more than two hypotheses were effectively tested in table 2.

Both the likelihood analysis in this paper and the pairwise comparison of Li (1993) used by Messier and Stewart (1997) assume that the d_N/d_S ratio is constant over all codon sites in the gene; that is, all amino acid sites in the lysozyme are under the same selectional pressure. This assumption must be highly unrealistic, as most amino acids in the lysozyme are conserved to maintain the structure and function of the protein (Stewart, Schilling, and Wilson 1987; Messier and Stewart 1997), and the intrinsic d_N/d_S ratios at these sites must be very low. The approaches of Messier and Stewart (1997) and of this paper detect positive selection only if the average d_N/d_S ratio over all sites in the lysozyme is significantly greater than one. This criterion is extremely stringent. Even when the average d_N/d_S ratio over the entire gene is considered, the assumption of a constant nonsynonymous rate among sites is expected to lead to underestimates of d_N, as nonsynonymous or amino acid substitution rates are variable among sites. For example, comparison of a gamma model of rates among sites (Yang 1994) and a constant-rate model using the amino acid sequences for the species of figure 1 leads to clear rejection of rate constancy ($P = 0.003$; the statistic is $2\Delta l = 2[-673.54 - (-677.84)] = 8.61$, and the estimated gamma parameter is 0.66).

Lastly, although the likelihood ratio test failed to reject the null hypothesis, $\omega_C \leq 1$, the alternative hypothesis, $\omega_C > 1$, indicating positive selection along lineage *c*, is never rejected. Little is known about the probability of type II errors associated with the test, that is, the probability of failing to reject the null hypothesis when it is wrong. Power analyses of this and other tests of positive selection appear to be very important. The likelihood ratio test established that the d_N/d_S ratio along branch *c* is significantly higher than the background ratio. This result is compatible with both relaxed selectional constraints and operation of positive selection along the lineage. Since lysozyme did not lose function along branch *c*, and, indeed, it acquired a new function, the hypothesis of positive selection seems at least as plausible as that of relaxed selectional constraint. To reach a decisive conclusion concerning the operation of positive selection along branch *c*, further studies seem necessary, especially those based on more realistic models of sequence evolution that account for nonsynonymous rate variation among sites.

Program Performance and Availability

The likelihood models developed in this paper are implemented in the codonml program in the PAML package (Yang 1997) and will be made available through the World Wide Web at http://abacus.gene.ucl.ac.uk/ziheng/paml.html. The analyses in this paper were performed on a DEC Alpha Station 500/433, which is about three times as fast as a Pentium Pro/200. Likelihood estimation under each model of table 1 took from a few hours to almost a day.

Acknowledgments

I thank Dr. Jim Leebens-Mack, two anonymous referees, and associate editors Dan Graur and Stanley Sawyer for criticisms and comments. I thank Dr. Masami Hasegawa of the Institute of Statistical Mathematics at Tokyo for his hospitality and for use of facilities during the revision of the manuscript.

Literature Cited

Collins, T. M., P. H. Wimberger, and G. J. P. Naylor. 1994. Compositional bias, character-state bias, and character-state reconstruction using parsimony. Syst. Biol. **43**:482–496.

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

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

Ina, Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. **40**:190–226.

Kendall, M., and A. Stuart. 1979. The advanced theory of statistics. Vol. 2, 4th edition. Charles Griffin and Company, London.

Li, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**:96–99.

Messier, W., and C.-B. Stewart. 1997. Episodic adaptive evolution of primate lysozymes. Nature **385**:151–154.

Maddison, W. P., and D. R. Maddison. 1992. MacClade: analysis of phylogeny and character evolution. Version 3. Sinauer, Sunderland, Mass.

Perna, N. T., and T. D. Kocher. 1995. Unequal base frequencies and the estimation of substitution rates. Mol. Biol. Evol. **12**:359–361.

Stewart, C.-B., J. W. Shilling, and A. C. Wilson. 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature **330**:401–404.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

———. 1997. PAML, a program package for phylogenetic analysis by maximum likelihood. CABIOS **13**:555–556.

Yang, Z., S. Kumar, and M. Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141**:1641–1650.

Stanley A. Sawyer, reviewing editor

Accepted February 10, 1998