Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages

Ziheng Yang and Rasmus Nielsen†*

*Galton Laboratory, Department of Biology, University College London; and †Department of Biometrics, Cornell University

The nonsynonymous (amino acid–altering) to synonymous (silent) substitution rate ratio ($\omega = d_N/d_S$) provides a measure of natural selection at the protein level, with $\omega = 1$, >1, and <1, indicating neutral evolution, purifying selection, and positive selection, respectively. Previous studies that used this measure to detect positive selection have often taken an approach of pairwise comparison, estimating substitution rates by averaging over all sites in the protein. As most amino acids in a functional protein are under structural and functional constraints and adaptive evolution probably affects only a few sites at a few time points, this approach of averaging rates over sites and over time has little power. Previously, we developed codon-based substitution models that allow the ω ratio to vary either among lineages or among sites. In this paper we extend previous models to allow the ω ratio to vary both among sites and among lineages and implement the new models in the likelihood framework. These models may be useful for identifying positive selection along prespecified lineages that affects only a few sites in the protein. We apply those branch-site models as well as previous branch- and site-specific models to three data sets: the lysozyme genes from primates, the tumor suppressor BRCA1 genes from primates, and the phytochrome (*PHY*) gene family in angiosperms. Positive selection is detected in the lysozyme and BRCA genes by both the new and the old models. However, only the new models detected positive selection acting on lineages after gene duplication in the *PHY* gene family. Additional tests on several data sets suggest that the new models may be useful in detecting positive selection after gene duplication in gene family evolution.

Introduction

The nonsynonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$) provides a sensitive measure of selective pressure at the amino acid level. An ω ratio greater than one means that nonsynonymous mutations offer fitness advantages and are fixed in the population at a higher rate than synonymous mutations. Positive selection can thus be detected by identifying cases where $\omega > 1$. Previous studies have most often employed a pairwise approach, calculating d_S and d_N rates between two sequences by averaging over all codons (amino acids) in the gene and over the time period that separates the sequences. Because many amino acids in a functional protein might be largely invariable (with ω close to 0) because of strong structural and functional constraints, the average d_N is rarely higher than the average d_S. As a result, this approach has little power in detecting positive selection (e.g., Endo, Ikeo, and Gojobori 1996; Sharp 1997; Akashi 1999; Crandall et al. 1999).

The model of codon substitution of Goldman and Yang (1994) (see also Muse and Gaut 1994) provides a framework for studying the mechanism of sequence evolution by comparing synonymous and nonsynonymous substitution rates. The original model assumes a single ω for all lineages and sites and has been extended to account for variation of ω either among lineages or among sites. The lineage-specific models (Yang 1998; Yang and Nielsen 1998; see also Muse and Gaut 1994) allow for variable ωs among lineages and are thus suitable for detecting positive selection along lineages. They assume no variation in ω among sites; as a result, they detect positive selection for a lineage only if the average d_N over all sites is higher than the average d_S. The site-specific models (Nielsen and Yang 1998; Yang et al. 2000) allow the ω ratio to vary among sites but not among lineages. Positive selection is detected at individual sites only if the average d_N over all lineages is higher than the average d_S. If adaptive evolution occurs at a few time points and affects a few amino acids (Gillespie 1991, pp. 132–139), both classes of models might lack power in detecting positive selection. It appears that averaging over sites is a more serious problem than averaging over lineages because the site-specific analysis has been successful in detecting positive selection in a variety of genes (Zanotto et al. 1999; Bishop, Dean, and Mitchell-Olds 2000; Yang et al. 2000; Fares et al. 2001; Haydon et al. 2001; Swanson et al. 2001). Computer simulations also confirmed the power of the site-specific analysis (Anisimova, Bielawski, and Yang 2001; see Yang and Bielawski 2000 for a review).

It appears worthwhile to develop models that allow the ω ratio to vary both among sites and among lineages. In this paper, we implement two such models. Our main objective is to improve the power of the likelihood ratio test (LRT) to detect positive selection along prespecified lineages. A major use of these new models might be to analyze the evolution of gene families, where functional divergence after gene duplication might have caused adaptive evolution (Ohta 1993). We implement the new models in the likelihood framework and apply them to analyze three data sets: the lysozyme genes from primates (Messier and Stewart 1997; Yang 1998), the cancer suppressor BRCA1 genes from primates (Huttley et al. 2000), and the phytochrome (*PHY*) gene family in angiosperms (Alba et al. 2001).

Key words: gene duplication, maximum likelihood, molecular adaptation, nonsynonymous substitution, positive selection, synonymous substitution.

Address for correspondence and reprints: Ziheng Yang, Department of Biology, 4 Stephenson Way, London NW1 2HE, UK.
E-mail: z.yang@ucl.ac.uk.

Mol. Biol. Evol. 19(6):908–917. 2002
© 2002 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

Table 1
Parameters in the Branch-Site Models

Site Class	Proportion	Background ω	Foreground ω
0	p_0	ω_0	ω_0
1	p_1	ω_1	ω_1
2	$p_2 = (1 - p_0 - p_1)p_0/(p_0 + p_1)$	ω_0	ω_2
3	$p_3 = (1 - p_0 - p_1)p_1/(p_0 + p_1)$	ω_1	ω_2

NOTE.—In model A, $\omega_0 = 0$ and $\omega_1 = 1$ are fixed, whereas in model B they are free to vary.

Theory

We assume that the phylogeny is known or independently estimated, and the branches that might be expected to be under positive selection are specified a priori. For example, in the analysis of a gene family, we are interested in testing whether positive selection has occurred along the lineages right after gene duplication. For convenience, we refer to branches for which we test positive selection as the ''foreground'' branches and all others the ''background'' branches.

The basic model of codon substitution specifies the substitution rate from sense codon i to sense codon j as

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

where κ is the transition-transversion rate ratio and π_j is the equilibrium frequency of codon j, calculated using the empirical nucleotide frequencies observed at the three codon positions (Goldman and Yang 1994). The scale factor μ is defined by the requirement that the average substitution rate is one:

$$-\sum_i \pi_i q_{ii} = 1. \quad (2)$$

Time and branch length are measured by the expected number of nucleotide substitutions per codon (Goldman and Yang 1994).

We assume that the ω ratio varies among codon (amino acid) sites, and there are four site classes in the sequence. The first class of sites are highly conserved in all lineages with a small ω ratio, ω_0. The second class includes neutral or weakly constrained sites at which $\omega = \omega_1$, where ω_1 is near or smaller than 1. In the third and fourth classes, the background lineages have ω_0 or ω_1, but the foreground branches have ω_2, which may be greater than 1. In other words, there are two site classes with the ratios ω_0 or ω_1 along the background branches, but along the lineages of interest, a certain event caused some sites to come under positive selection with the ratio $\omega_2 > 1$ (table 1). We assume that when positive selection occurs along the foreground lineages, it is equally likely to involve a site from site class 0 as a site from class 1; the proportions of sites from classes 2 and 3 are the same as those from classes 0 and 1 (table 1). This assumption can be relaxed by introducing an additional proportion parameter, but this is not pursued here.

We implement two versions of the model and refer to them later as models A and B. In model A, we fix $\omega_0 = 0$ and $\omega_1 = 1$. This model is an extension of the site-specific ''neutral'' model (M1) of Nielsen and Yang (1998), which assumes two site classes with $\omega_0 = 0$ and $\omega_1 = 1$ in all lineages. In model B, we estimate ω_0 and ω_1 from the data as free parameters. Although we envisage ω_0 and ω_1 to be smaller than one, we do not place this constraint in the implementation. Model B is an extension to the site-specific ''discrete'' model (M3) of Yang et al. (2000) with $K = 2$ site classes. In both models A and B, the proportions p_0 and p_1 as well as the ratio ω_2 are estimated from the data by maximum likelihood (ML).

Let the number of sites (codons) in the sequence be n and the observed data at site h be $\mathbf{x}_h$ ($h = 1, 2, \ldots, n$); $\mathbf{x}_h$ is a vector of codons at site h across all sequences in the alignment. Let y_h (=0, 1, 2, or 3) be the site class that site h belongs to. We assume that there are different classes of sites in the gene, but we do not know which class each site is from. Given the site class y_h, the conditional probability of observing data $\mathbf{x}_h$ at the site, $f(\mathbf{x}_h|y_h)$, can be calculated using previous algorithms. If the site is from classes 0 or 1 (if $y_h = 0$ or 1), all branches on the phylogeny have the same ω ratio, and $f(\mathbf{x}_h|y_h)$ can be calculated according to Goldman and Yang (1994). If the site is from classes 2 or 3 (if $y_h = 2$ or 3), the ω ratios are different for the background and foreground branches, and $f(\mathbf{x}_h|y_h)$ can be calculated according to Yang (1998). The unconditional probability is an average over the site classes:

$$f(\mathbf{x}_h) = \sum_{k=0}^{3} p_k f(\mathbf{x}_h|y_h = k). \quad (3)$$

We assume that the substitution process is independent among codon sites, and thus the log likelihood is a sum over all sites in the sequence

$$\ell = \sum_{h=1}^{n} \log\{f(\mathbf{x}_h)\}. \quad (4)$$

Parameters in the model, including branch lengths in the phylogeny, the transition-transversion rate ratio κ, as well as parameters in the ω distribution, are estimated by numerical maximization of the likelihood function (Yang 1997).

Models implemented here assume that the synonymous rate is constant across all sites, and only the nonsynonymous rate varies among site classes. The branch length (t), measured by the expected number of nucleotide substitutions per codon, is defined as an average across the site classes (Nielsen and Yang 1998). The scale factor μ in equation (1) differs between the foreground and background branches.

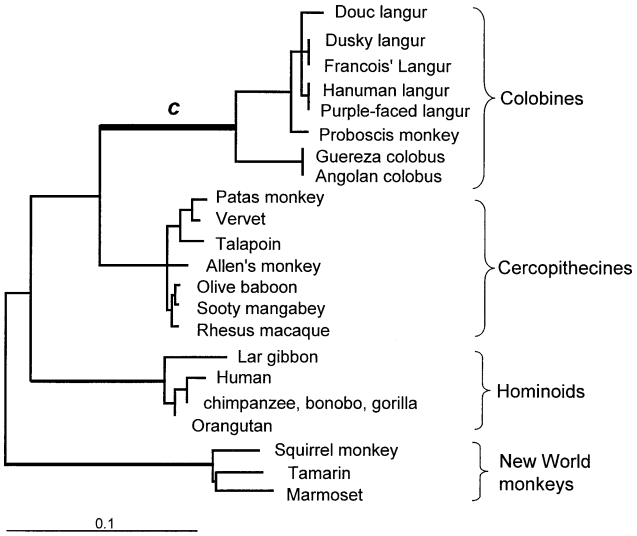

FIG. 1.—Phylogeny of 24 primate species for the lysozyme data set. Branch lengths, measured by the number of nucleotide substitutions per codon, are estimated under the model of codon substitution of Goldman and Yang (1994). Branch *c*, ancestral to the colobine monkeys, is the foreground branch for detecting positive selection.

After ML estimates of parameters are obtained, an empirical Bayes approach can be used to infer which class a site is most likely from (Nielsen and Yang 1998). The posterior probability that site h with data $\mathbf{x}_h$ is from site class k is

$$f(y_h = k \mid \mathbf{x}_h) = \frac{p_k f(\mathbf{x}_h \mid y_h = k)}{f(\mathbf{x}_h)} = \frac{p_k f(\mathbf{x}_h \mid y_h = k)}{\sum_j p_j f(\mathbf{x}_h \mid y_h = j)} \quad (5)$$

This approach does not account for sampling errors in the estimates of parameters. It is possible to use a hierarchical Bayes approach to accommodate uncertainties in parameter estimates by integrating over a prior distribution of parameters. The computation will be more complicated and can be achieved using Markov chain Monte Carlo. This approach is not pursued in this paper. We also note that parameter estimates obtained using other methods are applicable in the calculation of equation (5).

Real Data Analysis

Lysozyme Genes from Primates

The lysozyme *c* 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. Only the 19 distinct sequences are used, each with 130 codons. In many mammals, such as humans and rats, the lysozyme performs the function of fighting invading bacteria and exists mainly in secretions like tears and saliva as well as in white blood cells and tissue macrophages. Colobine monkeys (such as the langur) have fermentative foreguts, where high levels of lysozyme are present and where its function is to digest bacteria that pass from the foreguts into the true stomach (Stewart, Schilling, and Wilson 1987). Messier and Stewart (1997) suggested that positive selection occurred along the lineage ancestral to colobine monkeys (branch *c* in fig. 1). We apply the new models developed here to these data and treat branch *c* as the foreground branch and all other branches in the phylogeny as background branches (fig. 1).

Yang (1998) has performed a branch-specific likelihood analysis of the data, assuming that all sites in the sequence have the same ω ratio. The two-ratios model assigns the ratio ω_c for branch *c* and the ratio ω_0 for all other branches (table 2). This model fits the data significantly better than the one-ratio model of Goldman and Yang (1994). The LRT statistic for this comparison is $2\Delta\ell = 2 \times 2.13 = 4.26$, with $P = 0.039$ and df = 1 (table 2). So the ω ratio for branch *c* is significantly different from that for all other branches. To test whether ω_c is significantly higher than 1, the log likelihood value was calculated under the two-ratios model but with $\omega_c = 1$ fixed, giving the log likelihood value −1,042.50. The two-ratios model that does not place the constraint on ω_c (table 2) is not significantly better; the test statistic is $2\Delta\ell = 2 \times 1.33 = 2.66$, and $P = 0.10$ with df = 1. So ω_c is not significantly greater than 1 at the 5% significance level (see Yang 1998).

We also applied the site-specific likelihood models (Nielsen and Yang 1998; Yang et al. 2000) to the lysozyme data (table 2), which assume variable selective pressures among sites but no variation among branches in the phylogeny. We use three pairs of models, forming three LRTs: M1 (neutral) and M2 (selection), M0 (one-ratio) and M3 (discrete), and M7 (beta) and M8 (beta&ω) (Nielsen and Yang 1998; Yang et al. 2000). Model M1 (neutral) assumes two site classes with $\omega_0 = 0$ and $\omega_1 = 1$ fixed and with the proportions p_0 and p_1 estimated. Model M2 (selection) adds a third site class with the ratio ω_2 estimated. This model suggests that about 7% of sites are under positive selection with $\hat{\omega}_2 = 3.7$. Because model M2 (selection) is an extension of M1 (neutral), the two models can be compared using an LRT. The test statistic is $2\Delta\ell = 2 \times ((-1{,}035.83) - (-1{,}037.21)) = 2 \times 1.38 = 2.76$, with $P = 0.25$ and df = 2. So model M2 is not significantly better than M1. The discrete model (M3) with $K = 2$ site classes suggested that 18% of sites are under positive selection with $\hat{\omega}_1 = 2.6$ and identified six amino acid sites under positive selection at the 95% cutoff. Using $K = 3$ site classes produced the same estimates. M3 ($K = 2$) was significantly better than the one-ratio model; the test statistic is $2\Delta\ell = 17.20$ and $P < 0.001$ with df = 2. Model M7 (beta) assumes a beta distribution for ω over sites. The beta distribution is limited to the interval (0, 1) and provides a flexible null hypothesis for testing positive selection. The estimates suggest that the distribution reduced to the neutral model (M1). Model M8 (beta&ω) adds another site class to M7 (beta), with the ω ratio estimated from the data. The model suggested 16% of sites to be under positive selection with $\hat{\omega} = 2.5$ and identified seven sites under positive selection (the same six sites as under M3 plus site 17M). However, the difference between M7 and M8 is not statistically significant as $2\Delta\ell = 3.30$, with $P = 0.19$ and df = 2. We note

Table 2
Parameter Estimates for the Lysozyme Data

Model	p	ℓ	Estimates of Parameters	Positively Selected Sites
M0: one-ratio	1	−1,043.83	$\hat{\omega} = 0.574$	None
Branch-specific models (Model B in table 1 of Yang 1998)				
Two-ratios	2	−1,041.70	$\hat{\omega}_0 = 0.489$, $\hat{\omega}_0 = \mathbf{3.383}$	N/A
Site-specific models				
M1: neutral	1	−1,037.21	$\hat{p}_0 = 0.502$ ($\hat{p}_1 = 0.498$)	Not allowed
M2: selection	3	−1,035.83	$\hat{p}_0 = 0.498$, $\hat{p}_1 = 0.430$ ($\hat{p}_2 = \mathbf{0.072}$) $\hat{\omega}_2 = \mathbf{3.710}$	15L, 17M, 37G, 41R, 50R, 101R (at $0.5 < P < 0.8$)
M3: discrete ($K = 2$)	3	−1,035.23	$\hat{p}_0 = 0.823$ ($\hat{p}_1 = \mathbf{0.177}$) $\hat{\omega}_0 = 0.237$, $\hat{\omega}_1 = \mathbf{2.629}$	37G, 41R (at $P > 0.99$) 15L, 50R, 101R, 114N (at $P > 0.95$)
M3: discrete ($K = 3$)	5	Same at $K = 2$		
M7: beta	2	1,037.21	$\hat{p} = 0.011$, $\hat{q} = 0.011$	Not allowed
M8: beta&ω	4	1,035.56	$\hat{p}_0 = 0.788$, $\hat{p} = 99.65$, $\hat{q} = 298$ $\hat{p}_1 = \mathbf{0.212}$, $\hat{\omega} = \mathbf{2.538}$	37G, 41R (at $P > 0.99$) 15L 17M 50R 101R 114N (at $P > 0.95$)
Branch-site models				
Model A	3	−1,035.53	$\hat{p}_0 = 0.327$, $\hat{p}_1 = 0.269$ ($\hat{p}_2 + \hat{p}_3 = \mathbf{0.404}$) $\hat{\omega}_2 = \mathbf{4.809}$	Site for foreground lineage: 14R 21R 23I 87D (at $P > 0.9$) 41R 50R 126Q (at $P > 0.7$)
Model B	5	−1,034.27	$\hat{p}_0 = 0.611$, $\hat{p}_1 = \mathbf{0.157}$ ($\hat{p}_2 + \hat{p}_3 = \mathbf{0.232}$) $\hat{\omega}_0 = 0.166$, $\hat{\omega}_1 = \mathbf{2.319}$, $\hat{\omega}_2 = \mathbf{4.322}$	Sites for background ω_1: 15L 17M 37G 82S 101R 114N 125V ($0.7 < P < 0.8$) Sites for foreground ω_2: 14R 21R 23I 87D ($0.7 < P < 0.85$)

NOTE.—p is the number of free parameters for the ω ratios. Parameters indicating positive selection are presented in boldtype. Those in parentheses are presented for clarity only but are not free parameters; for example, under M8 (beta&ω), $p_1 = 1 - p_0$. Sites potentially under positive selection are identified using the human lysozyme sequence as the reference. Estimates of κ range from 4.1 to 4.6 among models.

that the M0-M3 comparison is more a test of variability in the ω ratio among sites and does not constitute a rigorous test of positive selection. The lysozyme gene has only 130 codons, and the nonsignificant results of the M1-M2 and M7-M8 comparisons might well be caused by the short sequence and lack of power of the LRTs. It is worth noting that parameter estimates under all the models, M2 (selection), M3 (discrete), and M8 (beta&ω), suggest presence of sites under positive selection.

The new branch-site models implemented in this paper are applied to the lysozyme data, with branch *c* of figure 1 considered as the foreground branch and all other branches in the tree as background branches. Model A does not allow for sites under positive selection across all lineages and suggests that a large proportion of sites (40%) are under positive selection along branch *c* with $\hat{\omega}_2 = 4.8$. This model can be compared with the site-specific model M1 (neutral); the LRT statistic is $2\Delta\ell = 2 \times 1.68 = 3.36$, with $P = 0.19$ and df = 2. So model A does not fit the data significantly better than model M1. Estimates under model B suggest sites under positive selection across all lineages ($\hat{\omega}_1 = 2.3$) as well as sites under selection along branch *c* only ($\hat{\omega}_2 = 4.3$). The comparison between model B and the site-specific model M3 (discrete with $K = 2$) gave $2\Delta\ell = 2 \times 0.96 = 1.92$, with $P = 0.38$ and df = 2. So model B does not fit the data significantly better than the site-specific model M3. However, this comparison is not interesting biologically because the null model allows for positive selection as well.

In sum, the selective pressure in the lysozyme is highly variable among sites. Some sites are under strong selective constraints throughout the primate phylogeny, whereas others appear to be under positive selection, with an excess of nonsynonymous substitutions relative to synonymous substitutions. The branch ancestral to the colobine monkeys (branch c in fig. 1) is under much stronger positive selection pressure. We note, however, that some of the LRTs fail to provide significant support for positive selection, possibly because of the short sequence and low divergence in the lysozyme data, resulting in low power in the tests.

The posterior probabilities that each site is from the four site classes of models A and B are calculated at the ML estimates of parameters (eq. 5). The probabilities for classes 2 and 3 are combined (see table 1). The sites identified in this way are listed in table 2. We also performed ML reconstruction of ancestral sequences (Yang, Kumar, and Nei 1995) using the codon model of Goldman and Yang (1994). The reconstruction suggested nine amino acid replacement changes along branch *c* in the tree: 14RK, 21RK, 23IV, 37DG, 41QE, 50QE, 62HR, 87DN, 126QK (see also Messier and Stewart 1997). These sites are essentially the same as those predicted to be under positive selection by the branch-site models (table 2).

Tumor Suppressor BRCA1 Gene from Primates

The tumor suppressor gene BRCA1 plays a role in the maintenance of genomic integrity, including recom-

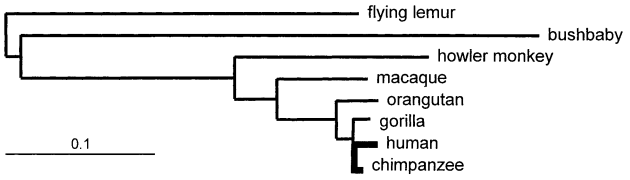

FIG. 2.—Phylogeny of primate species for the BRCA1 data set. The human and chimpanzee lineages are proposed to be under positive selection (Huttley et al. 2000).

binational and transcription-coupled DNA repair and in transcription regulation. Mutations in BRCA1 confer an increased risk of female breast cancer. The BRCA1 locus has a complex structure of 24 exons spanning more than 80 kb, with the majority (~60%) of the protein encoded by exon 11 (Huttley et al. 2000). Huttley et al. (2000) performed a lineage-specific ML analysis of the nucleotide sequences from exon 11 of human and nonhuman primates and suggested that the human and chimpanzee lineages are under positive selection (fig. 2). The authors hypothesized that the BRCA1 has a modified function in humans and chimpanzees relative to its homologues in other primates.

The alignment of Huttley et al. (2000) was modified slightly to accommodate the coding structure of the genes. The rat and mouse sequences used by the authors appear too divergent from the primate sequences, so that the alignment does not seem reliable in certain regions. Only the primate sequences are used in this paper. The alignment had 1,160 codons, but some regions had gaps, which are treated as ambiguity characters in the likelihood calculation (Yang 1997). The phylogenetic tree for the sequences is shown in figure 2.

The one-ratio model (M0) gives a log likelihood of −9,565.22, with the estimate $\hat{\omega} = 0.624$. The two-ratios model assigns two different ω ratios for the human and chimpanzee branches (ω_1) and for all other branches (ω_0). The log likelihood under this model is −9,561.06, with parameter estimates $\hat{\omega}_0 = 0.604$ for the background branches and $\hat{\omega}_1 = 2.676$ for the foreground branches. This model is significantly better than the one-ratio model ($2\Delta\ell = 2 \times 4.16 = 8.32$, $P = 0.0039$ and df = 1). To test whether ω_1 is significantly greater than 1, the two-ratios model is fitted to the data with $\omega_1 = 1$ fixed, giving a log-likelihood value of −9,562.72. This model is not significantly worse than the two-ratios model of table 2 without constraining $\omega_1 = 1$; the test statistic is $2\Delta\ell = 3.32$, with $P = 0.068$ with df = 1. Huttley et al. (2000) obtained a slightly larger test statistic, $2\Delta\ell = 4.3$, and their test was marginally significant ($P = 0.04$). This minor discrepancy seems to be caused by the different alignments.

We also applied the site-specific models (Nielsen and Yang 1998; Yang et al. 2000) to these data (table 3). The selective pressure on the protein varies greatly among amino acid sites. For example, using $K = 2$ site classes in the discrete model (M3) fits the data significantly better than the one-ratio model (M0); the test statistic is $2\Delta\ell = 458.64$, and $P = 0.000$ with df = 2. Model M3 suggests 17% of sites to be under positive selection with $\hat{\omega}_1 = 2.24$ and identifies seven amino acid sites under positive selection at the 95% cutoff (table 3). Model M8 (beta&ω) also suggests about 16% of sites under positive selection with $\hat{\omega} = 2.25$ and identifies the same seven sites under positive selection as model M3. Furthermore, M8 provides significantly better fit to the data than M7 ($2\Delta\ell = 15.24$, $P = 0.00049$, and df = 2). These tests provide significant evidence for presence of sites under positive selection. Unlike M3 (discrete) and M8 (beta&ω), model M2 (selection) does not suggest

Table 3
Parameter Estimates for the BRCA1 Gene

Model	p	ℓ	Estimates of Parameters	Positively Selected Sites
M0: one-ratio	1	−9,565.22	$\hat{\omega} = 0.624$	None
Branch-specific models (Yang 1998)				
Two-ratios	2	−9,561.06	$\hat{\omega}_0 = 0.604$, $\hat{\omega}_1 = 2.676$	
Site-specific models				
M1: neutral ($K = 2$)	1	−9,545.19	$p_0 = 0.290$, ($p_1 = 0.710$)	Not allowed
M2: selection ($K = 3$)	3	−9,542.06	$\hat{p}_0 = 0.000$, $\hat{p}_1 = 0.548$ $\hat{p}_2 = 0.451$), $\hat{\omega}_2 = 0.176$	None
M3: discrete ($K = 2$)	3	−9,535.90	$\hat{p}_0 = 0.834$ ($\hat{p}_1 =$ **0.166**) $\hat{\omega}_0 = 0.418$, $\hat{\omega}_1 =$ **2.240**	617H (at $P > 0.99$) 285P 479K 672G 892G 905Y 1144G (at $P > 0.95$)
M3: discrete ($K = 3$)	5	as above ($K = 2$)		
M7: beta	2	−9,543.52	$\hat{p} = 0.267$, $\hat{q} = 0.148$	Not allowed
M8: beta&ω	4	−9,535.90	$\hat{p}_0 = 0.836$, $\hat{p} = 71.8$, $\hat{q} = 99$ ($\hat{p}_1 =$ **0.164**), $\hat{\omega} =$ **2.249**	Same as under M3
Branch-site models				
Model A	3	−9,540.89	$\hat{p}_0 = 0.107$, $\hat{p}_1 = 0.244$ ($\hat{p}_2 + \hat{p}_3 =$ **0.649**), $\hat{\omega}_2 =$ **3.677**	See text
Model B	5	−9,533.13	$\hat{p}_0 = 0.636$, $\hat{p}_1 =$ **0.146** ($\hat{p}_2 + \hat{p}_3 =$ **0.218**) $\hat{\omega}_0 = 0.388$, $\hat{\omega}_1 =$ **2.086**, $\hat{\omega}_2 =$ **6.422**	See text

NOTE.—p is the number of free parameters for the ω ratios. Sites potentially under positive selection are identified using the human sequence as the reference. Estimates of κ range from 4.4 to 4.8 among models.

positive selection. As discussed by Yang et al. (2000), this pattern is because M1 (neutral) does not allow for sites with $0 < \omega < 1$; as a result, the extra site class in M2 (selection) is forced to account for such sites.

The branch-site models of this paper suggest sites under positive selection along the lineage of interest (table 3). Parameter estimates under model A suggest that 11% of sites are highly conserved across all lineages with $\hat{\omega}_0 = 0$, and 24% of sites are nearly neutral with $\omega_1 = 1$, whereas as high as 65% of sites are under strong positive selection along the human and chimpanzee branches with $\hat{\omega}_2 = 3.7$. Model A can be compared with the neutral model (M1) by an LRT. The statistic is $2\Delta\ell = 2 \times 4.30 = 8.60$, with $P = 0.014$ and df $= 2$. This improvement is statistically significant. Parameter estimates under model B (table 3) suggest that 15% of sites are under positive selection in all lineages with $\hat{\omega}_1 = 2.1$, whereas 22% of sites are under even stronger positive selection in the human and chimpanzee branches with $\hat{\omega}_2 = 6.4$. The LRT comparing the branch-site model B and the site-specific model M3 ($K = 2$) gave $2\Delta\ell = 2 \times 2.77 = 5.54$, and $P = 0.063$ with df $= 2$. This comparison is close to being significant. Because the null model in this comparison suggests positive selection at some sites along all lineages with $\omega_1 > 1$, it is safe to suggest positive selection along the human and chimpanzee lineages.

We examined the posterior probabilities for site classes under model A to infer which sites are likely to be under positive selection along the human and chimpanzee branches. Probabilities for site classes 2 and 3 were combined (see table 1). At the $P > 85\%$ level, the following sites are identified: 30A, 179E, 233I, 244K, 285P, 317T, 449G, 456A, 489K, 509K, 604N, 653P, 660A, 671S, 672G, 686C, 691E, 716G, 725N, 763R, 853I, 892G, 919Y, 975K, 1027N. We also applied ML reconstruction of ancestral sequences using the codon model of Goldman and Yang (1994) and Yang, Kumar, and Nei (1995). The reconstruction suggested the following amino acid changes along the human branch: 30TA, 179GE, 244RK, 285AP, 317MT, 449DG, 489NK, 604TN, 653LP, 660PA, 672RG, 716CG, 725YN, 763HR, 853VI, 919CY, 975RK, 1027SN and the following changes along the chimpanzee branch: 233IT, 456AV, 509KE, 671SC, 686CR, 691EQ, 892GE. The list identified by the branch-site model simply consists of the amino acids that probably changed along the foreground branches as suggested by the ancestral reconstruction.

Phytochrome Gene Family from Angiosperms

Light is of critical importance to plant metabolism and development, and plants employ a number of mechanisms to detect light. Phytochromes are the best-characterized plant photosensors. They are chromoproteins that regulate the expression of a large number of light-responsive genes and many photomorphogenic events, including seed germination, flowering, fruit ripening, stem elongation, and chloroplast development (Alba et al. 2001). All angiosperms characterized to date contain

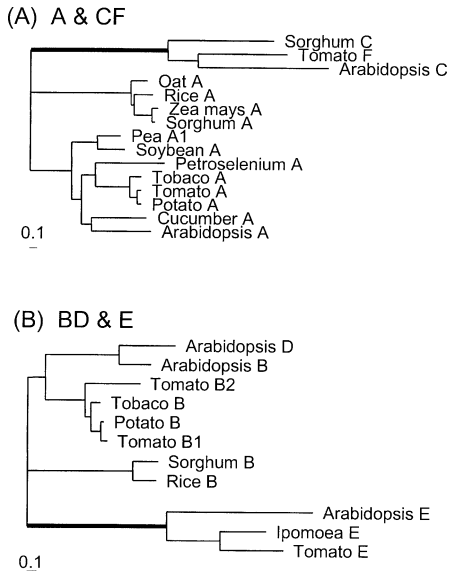

FIG. 3.—Phylogenies for the phytochrome (*phy*) gene family in angiosperms. (*A*) The branch separating the A and the C-F subfamilies is postulated to be under positive selection after the gene duplication. (*B*) The branch separating the B-D and E subfamilies is postulated to be under positive selection.

a small number of different PHY apoproteins encoded by a small *PHY* gene family. Alba et al. (2001) characterized the five loci in the *PHY* family in the tomato and performed phylogenetic analysis of the *PHY* family in angiosperms. The study characterized conserved and fast-evolving regions of the gene and demonstrated elevated evolutionary rates in between-subfamily comparisons. The authors pointed out that the rapid evolution is compatible with both positive selective pressure and relaxed evolutionary constraint. Here we apply the models of this paper to these data to detect possible positive selection driving the evolution of new functions after gene duplication. We use the alignment and phylogenetic tree published by Alba et al. (2001; fig. 4). To reduce the level of sequence divergence, we include the 16 sequences from the A and C-F subfamilies in one data set and the 11 sequences from the B-D and E subfamilies in another data set (see fig. 3*A* and *B*). Some sites in each data set had alignment gaps in all sequences, and we kept those sites to maintain the same site numbering (Alba et al. 2001). See Alba et al. (2001) for detailed sequence information and accession numbers. We test whether the gene was under positive selection along the branch separating the A and the C-F subfamilies (fig. 3*A*) and along the branch separating the B-D and E subfamilies (fig. 3*B*). Those branches are the foreground branches in our analysis. The results are reported in tables 4 and 5 for the ACF and BDE data sets, respectively.

The A and CF Subfamilies

The one-ratio model (M0) gives an estimate $\hat{\omega} = 0.089$ (table 4). This low average ratio suggests the dominating role of purifying selection in the evolution of the phytochrome gene family. The branch-specific

Table 4
Parameter Estimates for the Phytochrome Gene (AC&F)

Model	p	ℓ	Estimates of Parameters	Positively Selected Sites
M0: one-ratio	1	−29,984.12	$\hat{\omega} = 0.089$	None
Branch-specific models (Yang 1998)				
Two-ratios	2	−29,983.48	$\hat{\omega}_0 = 0.090$, $\hat{\omega}_1 = 0.016$	
Site-specific models				
M1: neutral ($K = 2$)	1	−31,641.62	$\hat{p}_0 = 0.291$, ($\hat{p}_1 = 0.709$)	Not allowed
M2: selection ($K = 3$)	3	−29,471.21	$\hat{p}_0 = 0.260$, $\hat{p}_1 = 0.051$ ($\hat{p}_2 = 0.689$), $\hat{\omega}_2 = 0.116$	None
M3: discrete ($K = 2$)	3	−29,413.70	$\hat{p}_0 = 0.608$ ($\hat{p}_1 = 0.392$) $\hat{\omega}_0 = 0.029$, $\hat{\omega}_1 = 0.211$	None
M3: discrete ($K = 3$)	5	−29,345.29	$\hat{p}_0 = 0.350$, $\hat{p}_1 = 0.498$ ($\hat{p}_2 = 0.151$) $\hat{\omega}_0 = 0.009$, $\hat{\omega}_1 = 0.097$, $\hat{\omega}_2 = 0.347$	None
M7: beta	2	−29,347.10	$\hat{p} = 0.630$, $\hat{q} = 5.191$	Not allowed
M8: beta&ω	4	−29,343.97	$\hat{p}_0 = 0.981$, $\hat{p} = 0.688$, $\hat{q} = 6.357$ ($\hat{p}_1 + 0.019$), $\hat{\omega} = 0.627$	None
Branch-site models				
Model A	3	−31,456.22	$\hat{p}_0 = 0.061$, $\hat{p}_1 = 0.139$ ($\hat{p}_2 + \hat{p}_3 = 0.800$), $\hat{\omega}_2 = 0.009$	None
Model B	5	−29,399.84	$\hat{p}_0 = 0.592$, $\hat{p}_1 = 0.348$ ($\hat{p}_2 + \hat{p}_3 =$ **0.060**) $\hat{\omega}_0 = 0.031$, $\hat{\omega}_1 = 0.221$, $\hat{\omega}_2 =$ **7.041**	105S 700A (at $P > 0.95$) 117P 227F 650T (at $P > 0.90$)

NOTE.—p is the number of free parameters for the ω distribution. Sites potentially under positive selection are identified using the *Zea mays* sequence as the reference. Estimates of κ range from 2.0 to 2.7.

two-ratios model assigns two ω ratios for the foreground (ω_1) and background branches (ω_0). Although the estimates $\hat{\omega}_1 > \hat{\omega}_0$ (table 4), the two-ratios model does not fit the data significantly better than the one-ratio model.

The site-specific models (Nielsen and Yang 1998; Yang et al. 2000) revealed a huge amount of variation in selective pressure among sites. For example, two site classes (M3, $K = 2$) fitted the data better than one site class (M0) by 570.42 log likelihood units. However, none of the site-specific models suggest the presence of sites under positive selection with $\omega > 1$, and all of them suggest that most sites in the sequence are under very strong selective constraints (table 4).

Branch-site model A adds an additional site class to the neutral model, where the selective pressure is different among the foreground and background branches. Although model A fits the data better than the neutral model, both models are highly unrealistic and even much worse than the one-ratio model (M0). There appears to be a large proportion of sites with ω small but

Table 5
Parameter Estimates for the Phytochrome Gene (BD&E)

Model	p	ℓ	Estimates of Parameters	Positively Selected Sites
M0: one-ratio	1	−22,361.51	$\hat{\omega} = 0.087$	None
Branch-specific models (Yang 1998)				
Two-ratios	2	−22,361.50	$\hat{\omega}_0 = 0.087$, $\hat{\omega}_1 = 0.085$	
Site specific models				
M1: neutral ($K = 2$)	1	−23,444.40	$\hat{p}_0 = 0.375$, ($\hat{p}_1 = 0.625$)	Not allowed
M2: selection ($K = 3$)	3	−22,049.53	$\hat{p}_0 = 0.328$, $\hat{p}_1 = 0.024$ ($\hat{p}_2 = 0.648$), $\hat{\omega}_2 = 0.131$	None
M3: discrete ($K = 2$)	3	−21,999.76	$\hat{p}_0 = 0.588$ ($\hat{p}_1 = 0.412$) $\hat{\omega}_0 = 0.022$, $\hat{\omega}_1 = 0.203$	None
M3: discrete ($K = 3$)	5	−21,982.67	$\hat{p}_0 = 0.397$, $\hat{p}_1 = 0.424$ ($\hat{p}_2 = 0.179$) $\hat{\omega}_0 = 0.009$, $\hat{\omega}_1 = 0.098$, $\hat{\omega}_2 = 0.304$	None
M7: beta	2	−21,983.55	$\hat{p} = 0.594$, $\hat{q} = 5.183$	Not allowed
M8: beta&ω	4	−21,983.55	As M7	
Branch-site models				
Model A	3	−23,281.87	$\hat{p}_0 = 0.085$, $\hat{p}_1 = 0.130$ ($\hat{p}_2 + \hat{p}_3 = 0.784$), $\omega_2 = 0.016$	None
Model B	5	−21,994.63	$\hat{p}_0 = 0.581$, $\hat{p}_1 = 0.388$ ($\hat{p}_2 + \hat{p}_3 =$ **0.031**) $\hat{\omega}_0 = 0.023$, $\hat{\omega}_1 = 0.207$, $\hat{\omega}_2 =$ **3.946**	902S (at ($P > 0.9$) 2S 287A 310G 446Y 535Q 640E 782I 851A 945S (at $P > 0.5$)

NOTE.—p is the number of free parameters for the ω ratios. Sites potentially under positive selection are identified using the sorghum B sequence as the reference; the numbering is the same as in table 4. Estimates of κ range from 2.0 to 2.6.

Table 6
The Effects of Codon Usage Bias on LRTs in the PHY Gene Family

Model	$2\Delta\ell$	Estimates Under M3 ($K = 2$)	Estimates Under Model B
ACF data set			
Fequal	214.56	$\hat{p}_0 = 0.573$, ($\hat{p}_1 = 0.427$) $\hat{\omega}_0 = 0.033$, $\hat{\omega}_1 = 0.247$	$\hat{p}_2 + \hat{p}_3 = 0.056$, $\hat{\omega}_2 = 7.37$
F3 × 4 (table 4)	27.72	$\hat{p}_0 = 0.608$, ($\hat{p}_1 = 0.392$) $\hat{\omega}_0 = 0.029$, $\hat{\omega}_1 = 0.211$	$\hat{p}_2 + \hat{p}_3 = 0.060$, $\hat{\omega}_2 = 7.04$
Fcodon.	22.44	$\hat{p}_0 = 0.605$, ($\hat{p}_1 = 0.395$) $\hat{\omega}_0 = 0.026$, $\hat{\omega}_1 = 0.192$	$\hat{p}_2 + \hat{p}_3 = 0.051$, $\hat{\omega}_2 = 24.05$
BDE data set			
Fequal	12.19	$\hat{p}_0 = 0.572$, ($\hat{p}_1 = 0.428$) $\hat{\omega}_0 = 0.027$, $\hat{\omega}_1 = 0.248$	$\hat{p}_2 + \hat{p}_3 = 0.034$, $\hat{\omega}_2 = 3.91$
F3 × 4 (table 5)	109.80	$\hat{p}_0 = 0.588$ ($\hat{p}_1 = 0.412$) $\hat{\omega}_0 = 0.022$, $\hat{\omega}_1 = 0.203$	$\hat{p}_2 + \hat{p}_3 = 0.031$, $\hat{\omega}_2 = 3.95$
Fcodon.	8.12	$\hat{p}_0 = 0.589$ ($\hat{p}_1 = 0.411$) $\hat{\omega}_0 = 0.020$, $\hat{\omega}_1 = 0.185$	$\hat{p}_2 + \hat{p}_3 = 0.027$, $\hat{\omega}_2 = 2.50$

NOTE.—All tests comparing site model M3 ($K = 2$) and site-branch model B are significant at the 1% level; df = 2.

positive, and neither model accounts for them. Branch-site model B is an extension to the discrete model with $K = 2$ site classes. The difference between those two models is significant, with $2\Delta\ell = 2 \times 13.86 = 27.72$, and $P = 10^{-6}$. Furthermore, parameter estimates under model B suggest a small proportion (6%) of sites under positive selection along the foreground branch with $\hat{\omega}_2 = 7.0$ (table 4). Therefore, model B detected positive selection along the foreground branch when the old models failed.

To examine how sensitive this result is to assumptions made in the model, we performed the LRT comparing M3 and model B under two assumptions about codon usage: the Fequal model, which assumes equal codon frequencies and thus ignores codon usage bias and the Fcodon model, which uses all the 61 codon frequencies as parameters. The results are shown in table 6. Previous studies suggested that codon usage bias can have a drastic effect on estimation of synonymous and nonsynonymous rates, much more important than the transition-transversion rate bias (see e.g., Yang and Bielawski 2000). It is thus interesting to examine the effect of codon usage bias on the LRT of positive selection. Results of table 6 suggest that although the LRT statistic varies according to the model assumption about codon usage, all of them are highly significant, supporting the presence of sites under positive selection among the foreground branch. The parameter estimates are quite similar and so are the calculated posterior probabilities (not shown). We also note that the two simpler models, Fequal and F3×4, are both rejected compared with Fcodon, indicating that codon usage is highly uneven.

ML reconstruction of ancestral sequences (Yang, Kumar, and Nei 1995) using the codon model of Goldman and Yang (1994) suggested as many as 277 amino acid substitutions along the foreground branch, which separates the A subfamily and the C-F subfamily. At the $P > 50\%$ level (eq. 5), the branch-site model B identifies the following 26 sites as potentially under positive selection along the branch: 52Q, 55R, 102T, 105S, 117P, 130T, 171T, 216E, 221T, 227F, 252I, 304I, 305D, 338V, 440L, 621N, 650T, 655S, 700A, 736K, 751G, 787N, 802V, 940T, 1069E, 1087D. Those are a subset of the sites predicted to have changed along the branch in ancestral reconstruction.

The BD and E Subfamilies

Results obtained from the B-D and E subfamilies are presented in table 5. The results are rather similar to those for the A and C-F subfamilies (table 4). For example, the overall ω ratio under model M0 (one-ratio) is 0.087, almost identical to the estimate from the ACF data set (0.089). The two-ratios model gave about the same fit to the data as the one-ratio model. The site-specific models indicated considerable variation in selective pressure among sites. For example, two site classes (M3, $K = 2$) fit the data better than one class (M0) by 638.25 log-likelihood units. However, none of the site-specific models suggested presence of sites under positive selection. Most sites appear to be under strong purifying selection. Branch-site model A is even much worse than the one-ratio model. Model B is much more realistic and fits the data better than M3 (discrete, $K = 2$), with $2\Delta\ell = 2 \times 54.90 = 109.8$, and $P < 0.000$, df $= 2$. Furthermore, parameter estimates under Model B suggested a small proportion of sites (3%) to be under positive selection on the foreground branch with $\hat{\omega}_2 = 3.9$. Hence, the new model B detected positive selection, whereas all old models failed. We also conducted this same test under two additional models of codon usage, Fequal and Fcodon, with the results shown in table 6. Although codon usage is highly biased, the parameter estimates are quite stable among those models, and the results of the LRT are robust to assumptions about codon usage bias.

We also applied the empirical Bayes procedure to calculate the posterior probabilities that each site comes from the four site classes, with the probabilities for site classes 2 and 3 combined (eq. 5, table 1). At the 50%

level, 10 codons are identified (table 5). For comparison, ML reconstruction of ancestral sequences under the codon model of Goldman and Yang (1994) suggested about 226 amino acid substitutions along the branch separating the B-D and the E subfamilies of the *PHY* gene. The 10 sites identified by model B are a subset of the sites predicted to have changed in ancestral reconstruction.

In sum, LRTs based on the branch-site model B confirmed the hypothesis that position selection was operating to drive functional divergence after gene duplication in both the ACF and the BDE data sets (fig. 3*A* and *B* and tables 4 and 5). This conclusion is also robust to assumptions about codon usage bias. However, we expect the prediction of sites under positive selection to be much less reliable (see *Discussion*). Sequencing more species to break the long branches in the phylogenies might provide information about which of the many amino acid changes were responsible for functional divergences.

Discussion

In the analysis of the lysozyme and the BRCA1 data sets, the previous branch and site models already detected positive selection, with different levels of statistical support. Although some of the LRTs were not significant in the lysozyme data set, we suspect that the nonsignificance may be because of the short sequence and lack of power of the tests. The phytochrome gene family is clearly under strong purifying selection, and previous models averaging rates either over branches or over sites failed to detect positive selection. However, the new branch-site model B detected positive selection causing adaptive changes at certain amino acid sites along branches right after the gene duplications that separated the A from the C-F subfamilies (fig. 3*A*) and that separated the B-D and the E subfamilies (fig. 3*B*).

The branch-site models have also been applied to several other gene families. In an analysis of the visual pigment gene family in vertebrates, the branch-site models detected positive selection along the lineage separating the rod and cone opsins, demonstrating the selective pressure exerted by the requirement of the new function of the ancestral rod opsin (B.S.W. Chang, personal communication). Bielawski and Yang (unpublished data) used the branch-site models to detect positive selection after gene duplication in the vertebrate Troponin *C* gene family, which created two distinct muscle isoforms for Troponin *C:* the fast skeletal muscle isoform and the cardiac and slow skeletal muscle isoform. A duplication of the caffeic acid *O*-methyl transferase gene apparently gave rise to the gene for isoeugenol *O*-methyl transferase, an enzyme that makes a floral scent compound important for attracting pollinators to flowers of the plant *Clarkia breweri* in the evening primrose family (Onagraceae). Todd Barkman (personal communication) applied the branch-site models to detect positive selection after that gene duplication event. In none of those data sets did previous branch and site models detect positive selection. Those case studies suggest that the new models have improved power in at least some data sets. Interestingly, all those examples involve gene family evolution.

In theory, the branch-site models implemented in this paper can be used to address two questions. The first is whether there are some sites in the gene that are under positive selection along the branches of interest; this is addressed by the LRT. The second is to identify positively selected sites when they exist. This is achieved using the Bayes prediction (eq. 5).

Real data analyses discussed earlier in the article suggest that the LRT based on the branch-site models has improved power in detecting positive selection than the branch models. Nevertheless, the new models make very specific assumptions, and it is advisable to examine the sensitivity of the LRTs to model assumptions and to sampling of sequences in the data. The models allow for only two types of lineages in the phylogeny. They will be difficult to apply if, for example, two branches are under positive selection but the selective pressure affects different amino acid sites along the two lineages. Fixing $\omega_0 = 0$ and $\omega_1 = 1$ in model A seems to lead to very poor fit in some data sets, especially when the gene is under strong purifying selection. Model B appears more widely applicable. We suggest that the null model M3 with only $K = 2$ site classes is not so restrictive as it might look. Although using two site classes with different ω ratios appears always to fit the data much better than the one-ratio model, it is seldom possible to fit more than three classes to real data; the parameter estimates simply collapse and the model reduces to one of fewer classes. Indeed use of three classes in M3 appears to fit any data set just as well as any parameter-rich continuous distributions (see Yang et al. 2000 for examples). Even use of three classes sometimes leads to collapse of the model into two classes, as in the lysozyme and BRCA1 data sets (tables 2 and 3). We also emphasize that the branches to be tested for positive selection should represent a priori biological hypotheses (Yang 1998). The significance values will not be correct if the LRT is applied to all branches in the phylogeny or if the same data are used to identify branches for test by the branch-site models. In particular, it is improper to use the ''free-ratios'' model of Yang (1998) to identify branches with high ω ratios and then use the branch-site models to test whether those branches are under positive selection.

Identifying amino acid residues under positive selection along the lineages of interest is clearly much more difficult than testing for the presence of such sites. We suspect that typical data sets might not contain enough information for the Bayes prediction or any other method to be reliable. Intuitively, the methods accumulate information about whether each site is under selection by comparing the numbers of synonymous and nonsynonymous substitutions at that site. In the site-specific analysis, many changes might have accumulated along branches of the phylogeny when many sequences are included in the alignment. The branch-site models, however, focus on only one or a few lineages of interest. If there is not enough opportunity for multiple changes at each site along those few lineages, the data will not

allow reliable estimation of parameters in the model, and calculation of the posterior probabilities using equation (5) will be affected. Indeed, Bayes prediction of sites does not seem to offer much advantage over simple ancestral sequence reconstruction. The two approaches returned identical or highly similar results in the three data sets analyzed in this paper.

Program Availability

The branch-site models are implemented in the codeml program in the PAML package (Yang 1997; http://abacus.gene.ucl.ac.uk/software/paml.html). ML estimation by numerical optimization is noted to be problematic for some of the data sets analyzed in this paper, and it is advisable to run the program multiple times to ensure convergence to the global optimum.

Acknowledgments

We thank Lee J. Pratt for kindly sending us the phytochrome gene alignment. Keith Crandall and two anonymous reviewers made many constructive comments. This study has been supported by UK Biotechnology and Biological Sciences Research Council grant 31/G14969 to Z.Y., US National Science Foundation Grant DEB-0089487 to R.N., and EU Human Frontier Science Program Grant RGY0055/2001-M to R.N. and Z.Y.

LITERATURE CITED

AKASHI, H. 1999. Within- and between-species DNA sequence variation and the 'footprint' of natural selection. Gene **238**: 39–51.

ALBA, R., P. M. KELMENSON, M.-M. CORDONNIER-PRATT, and L. H. PRATT. 2001. The phytochrome gene family in tomato and the rapid differential evolution of this family in angiosperms. Mol. Biol. Evol. **17**:362–373.

ANISIMOVA, M., J. P. BIELAWSKI, and Z. YANG. 2001. The accuracy and power of likelihood ratio tests to detect positive selection at amino acid sites. Mol. Biol. Evol. **18**:1585–1592.

BISHOP, J. G., A. M. DEAN, and T. MITCHELL-OLDS. 2000. Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. Proc. Natl. Acad. Sci. USA **97**:5322–5327.

CRANDALL, K. A., C. R. KELSEY, H. IMAMICHI, H. C. LANE, and N. P. SALZMAN. 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Mol. Biol. Evol. **16**: 372–382.

ENDO, T., K. IKEO, and T. GOJOBORI. 1996. Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. **13**:685–690.

FARES, M. A., A. MOYA, C. ESCARMIS, E. BARANOWSKI, E. DOMINGO, and E. BARRIO. 2001. Evidence for positive selection in the capsid protein-coding region of the foot-and-mouth disease virus (FMDV) subjected to experimental passage regimens. Mol. Biol. Evol. **18**:10–21.

GILLESPIE, J. H. 1991. The causes of molecular evolution. Oxford University Press, Oxford.

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

HAYDON, D. T., A. D. BASTOS, N. J. KNOWLES, and A. R. SAMUEL. 2001. Evidence for positive selection in foot-and-mouth-disease virus capsid genes from field isolates. Genetics **157**:7–15.

HUTTLEY, G. A., S. EASTEAL, M. C. SOUTHEY, A. TESORIERO, G. G. GILES, M. R. E. MCCREDIE, J. L. HOPPER, and D. J. VENTER. 2000. Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Nat. Genet. **25**:410–413.

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

MUSE, S. V., and B. S. GAUT. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. **11**:715–724.

NIELSEN, R., and Z. YANG. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929–936.

OHTA, T. 1993. Pattern of nucleotide substitution in growth hormone-prolactin gene family: a paradigm for evolution by gene duplication. Genetics **134**:1271–1276.

SHARP, P. M. 1997. In search of molecular Darwinism. Nature **385**:111–112.

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

SWANSON, W. J., Z. YANG, M. F. WOLFNER, and C. F. AQUADRO. 2001. Positive Darwinian selection in the evolution of mammalian female reproductive proteins. Proc. Natl. Acad. Sci. USA **98**:2509–2514.

YANG, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**: 555–556.

———. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. **15**:568–573.

YANG, Z., and J. P. BIELAWSKI. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. **15**:496–503.

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

YANG, Z., and R. NIELSEN. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. **46**:409–418.

YANG, Z., R. NIELSEN, N. GOLDMAN, and A.-M. K. PEDERSEN. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431–449.

ZANOTTO, P. M., E. G. KALLAS, R. F. SOUZA, and E. C. HOLMES. 1999. Genealogical evidence for positive selection in the *nef* gene of HIV-1. Genetics **153**:1077–1089.

KEITH CRANDALL, reviewing editor

Accepted February 5, 2002