CODON-SUBSTITUTION MODELS FOR DETECTING MOLECULAR ADAPTATION AT INDIVIDUAL SITES ALONG SPECIFIC LINEAGES

BU- 1567-M **May 2001**

Ziheng Yang
and
Rasmus Nielsen

Keywords: Gene duplication, Maximum likelihood, molecular adaptation, nonsynonymous substitution

Abstract:

The non-synonymous (amino acid-altering) to synonymous (silent) substitution rate ratio $(w = d_N/d_s)$ provides a measure of natural selection at the protein level, with w=1, < 1, and > 1 indicating neutral evolution, negative purifying selection, and positive diversifying selection, respectively. Previous studies that use 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 over sites and over time has little power. Previously we developed codon-based substitution models that allow the w ratio to vary either among lineages or among sites. In this paper we extend previous models to allow the w ratio to vary both among sites and among lineages and implement the new models in the likelihood framework. these models are useful for identifying positive selection along pre-specified lineages that affects only a few sites in the protein. The primate lysozyme and tumor suppressor BRCA1 genes were analyzed to evaluate the utility of the new methods. Positive selection is detected in both genes.

1567M

MBE, submitted
18 May 2001

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

Ziheng Yang[1] and **Rasmus Nielsen**[2]

[1]Galton Laboratory, Department of Biology, University College London, 4 Stephenson Way, London NW1 2HE, UK

[2]Department of Biometrics, 439 Warren Hall, Cornell University, Ithaca, NY 14853-7801, USA

Running Head: Codon Models of Molecular Adaptation

Keywords: Gene duplication, Maximum likelihood, Molecular adaptation, Nonsynonymous substitution, Positive selection, Synonymous substitution

Corresponding Author:

Dr Ziheng Yang
Department of Biology, 4 Stephenson Way, London NW1 2HE, UK
Phone: +44 (20) 7679 5083
Fax: +44 (20) 7383 2048
Email: z.yang@ucl.ac.uk

Abstract

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, negative purifying selection, and positive diversifying selection, respectively. Previous studies that use 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 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 are useful for identifying positive selection along pre-specified lineages that affects only a few sites in the protein. The primate lysozyme and tumour suppressor BRCA1 genes were analyzed to evaluate the utility of the new methods. Positive selection is detected in both genes.

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 synonymous (d_N) and nonsynonymous (d_S) rates between two sequences by averaging over all codons (amino acids) in the gene and over the time period that separates the sequences. As many amino acids in a functional protein might be largely invariable (with ω close to 0) due to 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., Sharp 1997; Endo, Ikeo and Gojobori 1996; 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 one 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) allow for variable ωs among lineages and are thus suitable for detecting positive selection along lineages. They assume no variation in ω among sites, and, as a result, 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), 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, as the site-specific analysis has been very successful in detecting positive selection in a variety of genes (Zanotto *et al.* 1999; Yang *et al.* 2000; Bishop,

Dean and Mitchell-Olds 2000; Haydon *et al.* 2001; Swanson *et al.* 2001; Fares *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 is 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 pre-specified lineages. A major use of those 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 two data sets: one of the lysozyme genes from primates (Messier and Stewart 1997; Yang 1998) and another of the cancer suppressor BRCA1 genes from primates (Huttley *et al.* 2000).

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 analysis of a gene family, we are interested in testing whether positive selection has occurred along the lineage 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; Muse and Gaut 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 then measured by the expected number of nucleotide substitutions per codon (Goldman and Yang 1994). The matrix of transition probabilities is given by

$$P(t) = \{p_{ij}(t)\} = e^{Qt}, \quad (3)$$

where $p_{ij}(t)$ is the probability that codon i will become codon j after time t. The calculation is accomplished by diagonalizing the rate matrix $Q = \{q_{ij}\}$ (Yang 1997).

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 branch has 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 become 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 to the site-specific "neutral" model 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. While 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 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. Note that 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:

$$f(\mathbf{x}_h) = \sum_{k=0}^{3} p_k f(\mathbf{x}_h \mid y_h = k). \tag{4}$$

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

$$\ell = \sum_{h=1}^{n} \log\{f(\mathbf{x}_h)\}. \tag{5}$$

Parameters in the model, including branch lengths in the phylogeny, the transition/transversion rate ratio κ, as well as any 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). Note that the scale factor μ in equation 2 differs between the foreground and background branches.

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

$$f(y_h = k \mid \mathbf{x}_h) = \frac{p_k f(\mathbf{x}_h \mid y_h = k)}{f(\mathbf{x}_h)} = \frac{p_k f(\mathbf{x}_h \mid y_h = k)}{\sum_j p_j f(\mathbf{x}_h \mid y_h = j)}. \quad (6)$$

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 6.

Real Data Analysis

Primate lysozyme evolution

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, 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 diversifying 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 d.f. = 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 d.f. = 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 likelihood ratio tests: 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 $\omega_2 = 3.7$. Since model M2 (selection) is an extension to 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 d.f. = 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 diversifying selection with $\omega_1 = 2.6$, 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 d.f. = 2. Model M7 (beta) assumes a beta distribution for ω over sites. The beta distribution is limited to the interval (0, 1) and so the model provides a flexible null hypothesis for testing positive selection. The estimates suggest that the distribution reduced to the neutral model (M1). Model M8 (beta&w) 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

$\omega = 2.5$, 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; the test statistic is $2\Delta\ell = 2 \times 1.65 = 3.30$, and $P = 0.19$, and d.f. = 2. Thus out of the three LRTs, only the one comparing M0 against M3 is significant. We note that the M0-M3 comparison is more a test of variability in the ω ratio among sites, and the M7 - M8 comparison is a stringent test of positive selection. A similar pattern was found in computer simulations, where the M0 - M3 comparison was significant much more often than the M7-M8 comparison (Anisimova, Bielawski and Yang 2001). While parameter estimates under all of models M2 (selection), M3 (discrete), and M8 (beta&ω) suggest presence of sites under positive selection, we suggest that the evidence be treated with caution, because not all the LRTs are significant.

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 suggest that a large proportion of sites (40%) are under positive selection along branch *c* with $\omega_2 = 4.8$. 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 d.f. = 2. So model A does not fit the data significantly better than model M1. Model B allows both for sites under positive selection across all lineages (if ω_0 or $\omega_1 > 1$) and for sites under selection along branch *c* only (if $\omega_2 > 1$). The estimates suggest existence of both classes of sites, that is, about 16% of sites under selection in all lineages with $\omega_1 = 2.3$ and about 23% of sites under even stronger selection along branch *c* with $\omega_2 = 4.3$. 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 d.f. = 2. So model B does not fit the data significantly better than the site-specific model M3. However, both models indicate presence of sites under positive selection along all lineages, and so the evidence for positive selection along branch *c* is stronger than indicated by this LRT. The branch-specific models are not nested within the new models, and so the simple LRT cannot be used to compare them.

In sum, the selective pressure in the lysozyme is highly variable among sites. There is evidence for positive selection affecting some sites throughout all lineages, and in particular, the lineage ancestral to the colobine monkeys appears to have a large proportion of sites under positive selection. However, most of the LRTs fail to provide significant support for positive selection. This result might be due to the short sequences and low divergences in the lysozyme data, resulting in lack of power in the LRTs.

Adaptive evolution in the tumour suppressor BRCA1 gene in primates

The BRCA1 plays a role in the maintenance of genomic integrity, including recombinational 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 80kb, 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 non-human primates, and suggested that the human and chimpanzee lineages are under positive diversifying 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 analyzed 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 $\omega = 0.624$. This is an average over all sites and all branches, and indicates that many sites are under purifying selection in the BRCA1 gene. The two-ratios model assigns two different ω ratios for the foreground human-chimpanzee branches (ω_1) and for all other branches (ω_0). The log likelihood under this model is –9,561.06, with parameter estimates $\omega_0 = 0.604$ for the background branches and $\omega_1 = 2.676$ for the

foreground branches. This model fits the data significantly better than the one-ratio model; the LRT statistic is $2\Delta\ell = 2 \times 4.16 = 8.32$, with $P = 0.0039$ with d.f. = 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 $2\Delta\ell = 2 \times 1.66 = 3.32$, with $P = 0.068$ with d.f. = 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 discrepancy seems to be due to the minor differences in the 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-raito model (M0); the test statistic is $2\Delta\ell = 2 \times ((-9{,}335.90) - (-9{,}565.22)) = 2 \times 229.32 = 458.64$, and $P = 0.000$ with d.f. = 2. Model M3 suggests 17% of sites to be under positive selection with $\omega_1 = 2.24$, 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 $\omega = 2.25$, and identifies the same seven sites under positive selection as model M3. Furthermore, M8 provides significantly better fit to the data than M7: the test statistic is $2\Delta\ell = 2 \times 6.62 = 15.24$, and $P = 0.00049$, with d.f. = 2. These tests provide significant evidence for presence of sites under diversifying selection. In contrast to models M3 (discrete) and M8 (beta&ω), model M2 (selection) does not suggest positive selection. As discussed by Yang et al. (2000), this pattern is because M1 (neutral) does not allow for sites with $0 < \omega < 1$, and 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 $\omega_0 = 0$, and 24% of sites are nearly neutral with $\omega_1 = 1$, while as high as 65% of sites are under strong positive selection along the human and chimpanzee branches with $\omega_2 = 3.7$. This

high proportion appears to be due to the fact that model A does not allow for sites under positive selection along all lineages. 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 d.f. = 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 $\omega_1 = 2.1$, while 22% of sites are under even stronger positive selection in the human and chimpanzee branches with $\omega_2 = 6.4$. 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 d.f. = 2. This comparison is close to being significant. Since both models suggest positive selection at some sites along all lineages with $\omega_1 > 1$, there is strong evidence that the human and chimpanzee branches are under diversifying selection.

We examined the posterior probabilities for site classes under model B to infer which sites are likely to be under positive selection along the human and chimpanzee branches. No site reached the 95% cutoff for any of site classes 1 (with ω_1), 2 (with ω_2), and 3 (with ω_2) (see table 1). Since both ω_1 and ω_2 are > 1, we combine the probabilities for those three site classes. Two sites have the combined $P > 99\%$: 617H and 1144G, and 14 more sites are identified at the 95% level: 179E, 285P, 317T, 384P, 471D, 479K, 509K, 670H, 672G, 676K, 684F, 892G, 905Y, 1027N.

Discussions

In the lineage-specific analysis, both the lysozyme and the BRCA1 genes show estimates of the ω ratio much larger than 1 for the lineages of interest. The site-specific analysis also suggested the presence of amino acids sites under positive selection in both proteins. Parameter estimates under the new branch-site models of this paper suggest much stronger positive selection along the lineages of interest in each gene. However, the results of the LRTs are mixed. In the lysozyme gene, we did not obtain statistically significant support for the new branch-site models over previous site-specific models. In the BRCA1 data set, we found that the new models fitted the data marginally significantly better than

the site-specific models, and there was significant evidence for presence of sites under positive selection. We note that when the site-specific model M3 already suggests sites under positive selection, the LRT comparing it against the new model B of this paper is not very interesting biologically.

We suspect that the new models of this paper might not often fit the data significantly better fit than previous site-specific models. Intuitively, the methods accumulate information about whether each site is under selection by comparing the numbers of synonymous and nonsynonymous substitutions at that site. For the simple site-specific models, many changes might have accumulated along branches of the phylogeny when many sequences are contained in the sequence alignment. The branch-site models, however, focus on only a few lineages of interest. If there is not enough opportunity for multiple changes at each site along these few lineages, the data will not contain sufficient information to reject the simple site- specific models.

The new models developed in this paper might be useful to analyze functional divergence after gene duplication. When a duplicated copy of a gene acquires a new function, the changed selective pressure might promote adaptive evolution by diversifying selection (Ohta 1993), which might affect only a few amino acids. Indeed, in an analysis of the evolution of the visual pigment family in vertebrates, neither the lineage-specific nor the site- specific analyses detect positive selection, but the new branch-site models detect 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, pers. comm.). So the new models have improved power in at least some data sets. Previous methods attempt to identify functional shifts by examining amino acid substitution rates along lineages of interest (Gu 1999). Such methods are expected to be less reliable than analysis based on codon-substitution models, as the amino acid substitution rate is not so sensitive a measure of selective pressure as is the d_N/d_S ratio.

LITERATURE CITED

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

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. in press.

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. U.S.A. **97:**5322-5327.

Crandall, K. A., C. R. Kelsey, H. Imamichi, H. C. Lane *et al.* 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 *et al.* 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.

Gu, X. 1999. Statistical methods for testing functional divergence after gene duplication. Mol. Biol. Evol. **16:**1664-1674.

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 *et al.* 2000. Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Nature 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. U.S.A. **98:**2509-2514.

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

Yang, Z. 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., 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.

Table 1

Parameters in the New 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

Table 2. Parameter estimates for the lysozyme data

Model	p	ℓ	Estimates of Parameters	Positively selected sites
M0: one-ratio	1	−1,043.83	$\omega = 0.574$	None
Branch-specific models (Model B in table 1 of Yang 1998)				
Two-ratios	2	−1,041.70	$\omega_0 = 0.489$, $\boldsymbol{\omega_c = 3.383}$	N/A
Site-specific models				
M1: neutral	1	−1,037.21	$p_0 = 0.502$ ($p_1 = 0.498$)	Not allowed
M2: selection	3	−1,035.83	$p_0 = 0.498$, $p_1 = 0.430$ ($\boldsymbol{p_2 = 0.072}$) $\boldsymbol{\omega_2 = 3.710}$	15L, 17M, 37G, 41R, 50R, 101R (at $0.5 < P < 0.8$)
M3: discrete ($K = 2$)	3	−1,035.23	$p_0 = 0.823$ ($\boldsymbol{p_1 = 0.177}$) $\omega_0 = 0.237$, $\boldsymbol{\omega_1 = 2.629}$	37G, 41R (at $P > .99$) 15L, 50R, 101R, 114N (at $P > .95$)
M3: discrete ($K = 3$)	5	Same as $K = 2$		
M7: beta	2	1,037.21	$p = 0.011$, $q = 0.011$	Not allowed
M8: beta&ω	4	1,035.56	$p_0 = 0.788$, $p = 99.65$, $q = 298$ $\boldsymbol{p_1 = 0.212}$, $\boldsymbol{\omega = 2.538}$	37G, 41R (at $P > .99$) 15L 17M 50R 101R 114N (at $P > .99$)
Branch-site models				
Model A	3	−1,035.53	$p_0 = 0.327$, $p_1 = 0.269$ ($\boldsymbol{p_2 = 0.404}$) $\boldsymbol{\omega_2 = 4.809}$	Sites for foreground lineage: 14R 21R 23I 87D (at $P > .9$) 41R 50R 126Q (at $P > .7$)
Model B	5	−1,034.27	$p_0 = 0.611$, $\boldsymbol{p_1 = 0.157}$ ($\boldsymbol{p_2 = 0.232}$) $\omega_0 = 0.166$, $\boldsymbol{\omega_1 = 2.319}$, $\boldsymbol{\omega_2 = 4.322}$	Sites for background ω_1: 15L 17M 37G 82S 101R 114N 125V ($.7 < P < .8$) Sites for foreground ω_2: 14R 21R 23I 87D ($.7 < P < .85$)

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.

Table 3. Parameter estimates for the BRCA1 gene

Model	p	ℓ	Estimates of Parameters	Positively selected sites
M0: one-ratio	1	–9,565.22	$\omega = 0.624$	None
Branch-specific models (Yang 1998)				
Two-ratios	2	–9,561.06	$\omega_0 = 0.604$, $\omega_1 = 2.676$	
Site-specific models				
M1: neutral ($K = 2$)	1	–9,545.19	$p_0 = 0.290$, ($p_1 = 0.710$)	Not allowed
			($\omega_0 = 0$, $\omega_1 = 1$)	
M2: selection ($K = 3$)	3	–9,542.06	$p_0 = 0.000$, $p_1 = 0.548$ ($p_2 = 0.451$)	None
			($\omega_0 = 0$, $\omega_1 = 1$), $\omega_2 = 0.176$	
M3: discrete ($K = 2$)	3	–9,535.90	$p_0 = 0.834$ (**$p_1 = 0.166$**)	285P 479K 672G 892G
			$\omega_0 = 0.418$, **$\omega_1 = 2.240$**	905Y 1144G (at $P > .95$)
				617H (at $P > .99$)
M3: discrete ($K = 3$)	5	as above ($K = 2$)		
M7: beta	2	–9,543.52	$p = 0.267$, $q = 0.148$	Not allowed
M8: beta&ω	4	–9,535.90	$p_0 = 0.836$, $p = 71.8$, $q = 99$	285P 479K 672G 892G
			(**$p_1 = 0.164$**), **$\omega = 2.249$**	905Y 1144G (at $P > .95$)
				617H (at $P > .99$)
Branch-site models				
Model A	3	–9,540.89	$p_0 = 0.107$, $p_1 = 0.244$ (**$p_2 = 0.649$**)	Many
			$\omega_0 = 0$, $\omega_1 = 1$, **$\omega_2 = 3.677$**	
Model B	5	–9,533.13	$p_0 = 0.636$, **$p_1 = 0.146$** (**$p_2 = 0.218$**)	Many
			$\omega_0 = 0.388$, **$\omega_1 = 2.086$, $\omega_2 = 6.422$**	

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.

Figure legends

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 columbine monkeys, is the foreground branch for detecting positive selection.

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).