J Mol Evol (2000) 51:423–432
DOI: 10.1007/s002390010105

JOURNAL OF MOLECULAR EVOLUTION

Maximum Likelihood Estimation on Large Phylogenies and Analysis of Adaptive Evolution in Human Influenza Virus A

Ziheng Yang

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

Received: 25 April 2000 / Accepted: 24 July 2000

Abstract. Algorithmic details to obtain maximum likelihood estimates of parameters on a large phylogeny are discussed. On a large tree, an efficient approach is to optimize branch lengths one at a time while updating parameters in the substitution model simultaneously. Codon substitution models that allow for variable nonsynonymous/synonymous rate ratios ($\omega = d_N/d_S$) among sites are used to analyze a data set of human influenza virus type A hemagglutinin (HA) genes. The data set has 349 sequences. Methods for obtaining approximate estimates of branch lengths for codon models are explored, and the estimates are used to test for positive selection and to identify sites under selection. Compared with results obtained from the exact method estimating all parameters by maximum likelihood, the approximate methods produced reliable results. The analysis identified a number of sites in the viral gene under diversifying Darwinian selection and demonstrated the importance of including many sequences in the data in detecting positive selection at individual sites.

Key words: Large phylogenies — Maximum likelihood — Molecular adaptation — Nonsynonymous substitution — Phylogenetics — Positive selection — Synonymous substitution

Introduction

The nonsynonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$) provides a sensitive measure of selective pressure at the protein level and is particularly useful for identifying adaptive protein evolution. Values of $\omega = 1$, < 1, and > 1 indicate neutral evolution, purifying (negative) selection, and diversifying (positive) selection on the protein, respectively. Early studies estimate synonymous (d_S) and nonsynonymous (d_N) substitution rates by averaging over all sites in the protein. As many amino acids may be largely invariable due to functional constraints with ω close to 0 and adaptive evolution most likely affects only a few amino acids, such analysis rarely find ω ratios >1 or detect positive selection (Crandall et al. 1999). Recently, methods have been developed that account for variable selective pressures among sites. Fitch et al. (1997; see also Bush et al. 1999) and Suzuki and Gojobori (1999) inferred sites under positive selection by reconstructing ancestral sequences using parsimony and counting synonymous and nonsynonymous changes along the tree at each site. Maximum likelihood (ML) methods based on explicit models of codon substitution assuming variable ω ratios among sites were developed by Nielsen and Yang (1998) and Yang et al. (2000). Those methods have been found to be powerful in detecting adaptive evolution at a few sites in a background of purifying selection (e.g., Fitch et al. 1997; Nielsen and Yang 1998; Bush et al. 1999; Suzuki and Gojobori 1999; Bishop et al. 2000; Yang et al. 2000). For example, in a recent analysis of the *nef* gene in HIV-1, Zanotto et al. (1999) found that the ML method of Nielsen and Yang (1998), which accounts for variable ω ratios among sites, detected a number of sites under positive selection, while both pairwise comparison and sliding window analysis, which average synonymous and nonsynonymous rates over the gene or gene segment, failed.

The ML method has several advantages. It has a

sound statistical basis and accounts for uncertainties in unknown ancestral sequences. The substitution models used in ML account for different transition and transversion rates and biased codon usage, important features of DNA sequence evolution often ignored by other methods. The likelihood method thus provides a powerful framework for testing for the presence of sites under positive selection and for identifying them. However, the ML method involves intensive computation, especially for large data sets. As including more sequences in the data increases the numbers of synonymous and nonsynonymous changes at each site along the tree and thus improves the power to detect positive selection, it is important to improve ML algorithms so that the method can be used to analyze large data sets.

In this paper, I discuss computational issues and algorithmic details of ML parameter estimation on a large phylogeny to stimulate development of efficient ML algorithms. As much of the computation is spent on estimation of branch lengths by numerical optimization, I also explore the possibility of using approximate methods to estimate branch lengths for codon models and then using them to test for selection and to infer sites under selection. The data set of human influenza virus A hemagglutinin (HA) gene, previously analyzed by Bush et al. (1999), is used as a test data set; it has 349 distinct sequences.

Maximum Likelihood Estimation on a Large Phylogeny

Estimation of Branch Lengths Under Site-Homogeneous Models

On a large phylogeny, great saving can be achieved by optimizing branch lengths one by one. This idea has been used in programs such as MOLPHY (Adachi and Hasegawa 1996), PAUP* (Swofford 1999), and PHYLIP (Felsenstein 1993). In this section I discuss this algorithm for estimating branch lengths with other parameters (that is, those in the substitution model) fixed. Models of codon substitution are used as examples, although the algorithm works for nucleotide- and amino acid–based analyses as well.

I will first describe the algorithm for site-homogeneous models, which were developed by Felsenstein (1981) for nucleotides, Kishino et al. (1990) for amino acids, and Goldman and Yang (1994) and Muse and Gaut (1994) for codons. The models assume independence of data among sites. Let the data at any site h be $\mathbf{x}_h$; for the tree of Fig. 1, $\mathbf{x}_h$ is the set of codons observed in all the 18 sequences at site h. The likelihood, that is, the probability of observing the entire sequence data set, is the product of the probabilities of observing data at individual sites. The log likelihood is thus a sum over sites in the sequence.

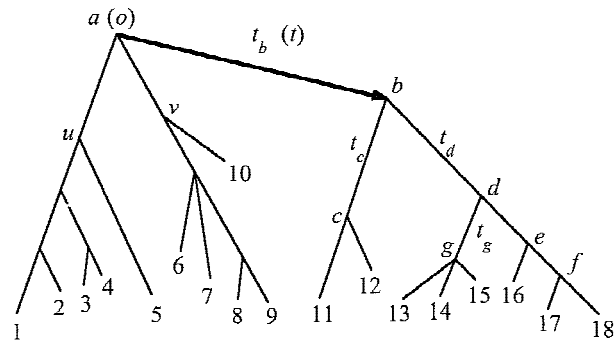

Fig. 1. A phylogeny used to explain the likelihood calculation.

$$\ell = \sum_{h=1}^{n} \log\{f_h\}, \tag{1}$$

where n is the number of sites in the sequence, and $f_h = f(\mathbf{x}_h)$ is the probability of observing data $\mathbf{x}_h$ at site h. Note that if two sites have the same data (site pattern), f will be the same and will be calculated only once. The likelihood calculation is thus proportional to the number of distinct site patterns. Below, I concentrate on calculation of f_h for one particular site h.

Probability f_h is efficiently calculated using the pruning algorithm of Felsenstein (1981). The models discussed here are time-reversible and cannot identify the root of the tree. The root can thus be placed at any place on the unrooted tree to simplify calculation. Consider estimation of the branch length t_b (or t) for branch a–b in Fig. 1. We place the root o at node a but consider it to be ancestral to a and b; that is, node o has daughter nodes a and b while node a has daughter nodes u and v (Fig. 1). We use t_i to denote the length of the branch leading to node i, and use x_i to denote the character (nucleotide, amino acid, or codon) at node i at that site; x_i is observed if node i is a tip of the tree and is unknown if node i is an ancestral node.

Let $L_i(x_i)$ be the probability of observing data at the site at the tips of the tree that are descendents of node i, given that node i has character x_i. For example, $L_d(x_d)$ is the conditional probability of observing characters in species 13–18 at the site, given that node d has character x_d (see Fig. 1). This was termed the "conditional likelihood" in Felsenstein (1981). If node i is a tip, $L_i(x_i) = 1$ if x_i is the observed character and 0 otherwise. If the data contain unidentified nucleotides, $L_i(x_i)$ is set to 1 for each x_i that is compatible with the ambiguity data (J. Felsenstein personal communication). For example, if the data at the site is codon TTR, where R refers to purine (A or G), then $L_i(x_i) = 1$ if x_i is TTA or TTG and $= 0$ if x_i is any other codon. Alignment gaps are either removed or treated as ambiguity characters; both approaches under-

estimate sequence divergences. If node i has two daughter nodes j and k, we have

$$L_i(x_i) = \left[\sum_{x_j} p_{x_i x_j}(t_j) L_j(x_j)\right] \times \left[\sum_{x_k} p_{x_i x_k}(t_k) L_k(x_k)\right]. \quad (2)$$

The product involves as many terms as the number of daughter nodes of node i. Thus the conditional probabilities $L_i(x_i)$ are calculated for tips and daughter nodes before ancestral nodes, with $L_o(x_o)$ for the root calculated last. This is the famous "pruning" algorithm (Felsenstein 1981). Since $L_o(x_o)$ is the probability, conditional on x_o, of observing data (at the site) in all sequences, the unconditional probability is

$$f_h = \sum_{x_o} \pi_{x_o} L_o(x_o) = \sum_{x_a}\sum_{x_b} \pi_{x_a} p_{x_a x_b}(t) L_a(x_a) L_b(x_b). \quad (3)$$

This is obtained by using Eq. 2—with o, a, and b replacing i, j, and k, respectively—and noting that $p_{x_o x_a}(0) = 1$ if $x_o = x_a$ and 0 otherwise. To calculate the derivatives of ℓ with respect to branch length t, note that the transition probability from codon x to codon y over time t is

$$p_{xy}(t) = \sum_k c_{xyk} e^{\lambda_k t}, \quad (4)$$

where λ_k and c_{ijk} are functions of the substitution rate matrix, independent of t (e.g., Grimmett and Stirzaker 1992, p. 242; Yang and Kumar 1996). Thus

$$p'_{xy}(t) = \frac{\partial p_{xy}(t)}{\partial t} = \sum_k c_{xyk}\lambda_k e^{\lambda_k t},$$
$$p''_{xy}(t) = \frac{\partial^2 p_{xy}(t)}{\partial t^2} = \sum_k c_{xyk}\lambda_k^2 e^{\lambda_k t}. \quad (5)$$

Since $L_a(x_a)$ and $L_b(x_b)$ in Eq. 3 are free of t, we have

$$f' = \frac{\partial f_h}{\partial t} = \sum_{x_a}\sum_{x_b} \pi_a p'_{x_a x_b}(t) L_a(x_a) L_b(x_b),$$
$$f'' = \frac{\partial^2 f_h}{\partial t^2} = \sum_{x_a}\sum_{x_b} \pi_a p''_{x_a x_b}(t) L_a(x_a) L_b(x_b). \quad (6)$$

Finally, we have from Eq. 1

$$\ell' = \frac{\partial \ell}{\partial t} = \sum_h \frac{f'}{f}$$
$$\ell'' = \frac{\partial^2 \ell}{\partial t^2} = \sum_h \frac{f f'' - (f')^2}{f^2}. \quad (7)$$

The availability of the second derivatives allows us to use an efficient modified Newton method, which updates t according to the following formula (e.g., Gill et al. 1981):

$$t^{(k+1)} = t^{(k)} + \alpha \ell'/\ell''. \quad (8)$$

The Newton method has step length $\alpha = 1$, but sometimes diverges. Thus the following modification is made (Gill et al. 1981). If the likelihood $\ell^{(k+1)}$ at $t^{(k+1)}$ is worse than the old value $\ell^{(k)}$ at $t^{(k)}$, we reduce the step length α, say, by halving it repetitively, until the new value is not worse. This modified Newton algorithm is nondecreasing.

As $L_a(x_a)$ and $L_b(x_b)$ are independent of t, they do not need to be recalculated when t is updated using Eq. 8, and thus calculation of the log likelihood and its derivatives (ℓ, ℓ', ℓ'') is fast. To estimate another branch length, we move the root to the new branch. For example, to estimate t_g in Fig. 1, we place the root on branch d–g with zero distance to node g. Then conditional probabilities for nodes on the path from the new root to the old root (that is, nodes d and b) have to be updated; those for other nodes are unchanged. Branch lengths are optimized in this way one by one. A cycle is completed after all branch lengths are optimized. As estimates of branch lengths are correlated, several cycles are needed to achieve convergence of all branch lengths in the tree.

Estimation of Branch Lengths Under Site-Class Models

Site-class models refer to models that assume a statistical distribution (several site classes) to account for the heterogeneity of the substitution process among sites. Examples include the codon-based models that account for different selective pressures indicated by the ω ratio (Nielsen and Yang 1998; Yang et al. 2000) and the gamma model of variable rates among nucleotide or amino acid sites (Yang 1994). Under such models, the probability of observing data at a site, f_h, is an average over the site classes. For example, the codon model M3 (discrete, Yang et al. 2000) assumes a general discrete distribution for ω, so that the sequence has K classes of sites, in proportions $p_0, p_1, \ldots, p_{K-1}$ and with ω ratios $\omega_0, \omega_1 \ldots, \omega_{K-1}$. The conditional probability of data $\mathbf{x}_h$ given that the site is from class k, $f(\mathbf{x}_h|\omega_k)$, is calculated in the same way as under the site-homogeneous model (Goldman and Yang 1994). The transition probability $p_{xy}(t)$ will also depend on the site class, and so we write it as $p_{xy}(t;\omega_k)$. As we do not know which class the site is from, the unconditional probability is an average over the ω distribution

$$f_h = f(\mathbf{x}_h) = \sum_k p_k f(\mathbf{x}_h|\omega_k) \quad (9)$$

(Nielsen and Yang 1998; Yang et al. 2000).

To estimate branch length t_b (or t) in Fig. 1, we have, from Eq. 3,

$$f_h = \sum_k \sum_{x_a} \sum_{x_b} p_k \pi_a p_{x_a x_b}(t;\omega_k) L_a(x_a;\omega_k) L_b(x_b;\omega_k). \quad (10)$$

Similarly, the derivatives are

$$f' = \frac{\partial f_h}{\partial t} = \sum_k \sum_{x_a} \sum_{x_b} p_k \pi_a p'_{x_a x_b}(t;\omega_k) L_a(x_a;\omega_k) L_b(x_b;\omega_k),$$

$$f'' = \frac{\partial^2 f_h}{\partial t^2} = \sum_k \sum_{x_a} \sum_{x_b} p_k \pi_a p''_{x_a x_b}(t;\omega_k) L_a(x_a;\omega_k) L_b(x_b;\omega_k). \quad (11)$$

The conditional probabilities $L_i(x_i)$ are now calculated for each site class (that is, for each ω_k). Apart from calculations of f, f', and f'', the algorithm described above for site-homogeneous models applies.

It may be noted that the site-homogeneous algorithm is used for models that assume different substitution parameters for prior partitions of sites in the sequence. Examples include models of Yang (1996b), which assume different rates and transition/transversion rate ratios for the three codon positions. As we know a priori which codon position each site is from, those models are computationally different from the site-class models. It should also be noted that the algorithms for estimating branch lengths one at a time works only when the molecular clock (rate constancy among lineages) is not assumed.

Memory Requirement

The algorithm discussed above works efficiently if the conditional probabilities $L_i(x_i)$ are stored in the computer memory. For the site-homogeneous models,

$$p \times c \times d \times 8 \quad (12)$$

bytes of space are needed, where p is the number of site patterns, c is the number of character states (4 for nucleotides, 20 for amino acids, and 61 for codons under the universal genetic code), d is the number of nodes, and 8 is the size of a double number (8 bytes on most systems). A bifurcating tree with s species has $d = 2s - 2$ nodes. As discussed above, if the data do not contain ambiguity characters or alignment gaps, $L_i(x_i)$ for tips will be either 0 or 1 and are not stored in memory, in which case $d = s - 2$.

Under the site-class models, the conditional probabilities $L_i(x_i; \omega_k)$ have to be stored for each site class k, so that K times as much space is needed to store the conditional probabilities for internal nodes of the tree. Note that $L_i(x_i)$ for tips do not depend on the site class k.

Scaling to Avoid Underflows

On a larger phylogeny, the conditional probabilities $L_i(x_i)$ or $L_i(x_i; \omega_k)$ can easily become too small to represent in the computer. Such underflows can be avoided by dividing the $L_i(x_i)$'s for different x_i by a very small scale factor and by adding its logarithm to $\log\{f_h\}$ (Eq. 1) at the end of the calculation of the log likelihood. For example, to perform scaling at node i, we find the maximum of $L_i(x_i)$ for different characters x_i. Let this be L_m, which may be a very small number. Then we divide each $L_i(x_i)$ by L_m and the calculation proceeds as usual. At the end of the likelihood calculation, we add $\log\{L_m\}$ to $\log\{f_h\}$ (Eq. 1). Scaling is performed for the chosen nodes for each site h in the sequence. It appears sufficient to perform scaling for every 50 or 100 descendent nodes visited during the pruning algorithm. Under site-class models, scaling is performed separately for the K site classes.

The algorithms of updating branch lengths one at a time requires the logarithms of scaling factors, $\log\{L_m\}$, to be stored in memory. The space required is $p \times d_s \times 8$ bytes for site-homogeneous models, where d_s is the number of nodes chosen for scaling (about $s/50$, say), and K times as much for site-class models. This memory demand is trivial.

Estimation of Substitution Parameters

Parameters in the substitution models, such as the transition/transversion rate ratio κ and the d_N/d_S rate ratio ω, can be estimated using any of the standard nonlinear programming algorithms, updating all parameters simultaneously. The commonly used conjugate gradient and quasi-Newton methods make use of first derivatives, which can be approximated using the difference method. The BFGS algorithm (Gill et al. 1981) is used in PAML (Yang 1997). Methods that do not use derivatives such as Powell's method (e.g., Brent 1973) are also usable. For large phylogenies, the following algorithm appears feasible, and it cycles through two phases. In phase I, substitution parameters are updated simultaneously with an algorithm like BFGS while branch lengths are fixed. In phase II, branch lengths are updated one by one, while substitution parameters are fixed. The procedure has to be repeated to achieve global convergence of all parameters. It is noted that optimization of substitution parameters (phase I) takes more time than estimating branch lengths (phase II). Thus in early stages of the algorithm, substitution parameters are optimized only crudely. This algorithm is used to analyze the large data set of this paper.

Another algorithm, used in early versions of PAML (Yang 1997), is to update all parameters including branch lengths simultaneously using the BFGS algorithm, with first derivatives calculated using the difference approximation. The relative performance of the two

algorithms can be very different and depends on many factors. If most parameters to be estimated are branch lengths as in a large phylogeny, or if branch lengths only are estimated, the algorithm of updating one branch length at a time is more efficient. Examples include the HKY85 substitution model with κ fixed or the gamma model of rates for sites (Yang 1997) with the shape parameter fixed. If the data set is small (say, with fewer than 10 or 20 sequences) and substitution parameters need to be estimated, the algorithm of simultaneous updating may be more efficient. When branch lengths and substitution parameters are highly correlated, as in models of variable substitution rates among sites (Yang 1996a), the algorithm of updating one branch length at a time can be very inefficient.

Approximate Branch Lengths Under Models of Codon Substitution

The feasibility of using approximate branch lengths in codon-based likelihood analysis of adaptive evolution is tested. The branch length used in codon substitution models is defined as the expected number of nucleotide substitutions per codon (Goldman and Yang 1994). This can be approximated in a number of ways. For example, several methods exist to estimate the numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions per site. The number of nucleotide substitutions per codon is thus

$$t = 3 \times (Sd_S + Nd_N)/(S + N), \quad (13)$$

where S and N are the numbers of synonymous and nonsynonymous sites in the sequence, respectively (Goldman and Yang 1994; Yang and Nielsen 1998). Note that $S/(S + N)$ is the proportion of synonymous sites and can be calculated easily for any substitution model (Goldman and Yang 1994; Ina 1995). Branch lengths can then be estimated by the least-squares method. In this paper, I test the option of using the method of Nei and Gojobori (1986, NG) to estimate d_S and d_N, and then using neighbor-joining (Saitou and Nei 1987) or least squares to estimate branch lengths. This option is referred to later as approximate method I. It is not expected to match the codon models closely, as the latter account for biased transition and transversion rates and biased codon usage, which are ignored by NG. Methods for estimating d_S and d_N that account for those features (Goldman and Yang 1994; Yang and Nielsen 2000) may produce results closer to estimates from the codon models.

Another option tested in this paper is to use a nucleotide-based likelihood analysis to estimate branch lengths for codon-substitution models. Although the same data are used, the nucleotide-based analysis requires much less computation, as the matrices are of size 4 × 4 rather than of 61 × 61. The branch length in a nucleotide substitution model is conventionally defined as the expected number of nucleotide substitutions per (nucleotide) site. Thus the branch length in the codon model is approximated by the sum of branch lengths at the three codon positions. Here I use a nucleotide model of Yang (1996b) that accounts for different substitution rates, base frequencies, and transition/transversion rate ratios at the three codon positions (BASEML, Mgene = 4 in PAML). This model is very close to the codon model of Goldman and Yang (1994). The method is referred to as approximate method II.

Analysis of Human Influenza Virus Type A Hemagglutinin Gene

Sequence Data and Analysis

The human influenza virus type A hemagglutinin (HA) gene from 357 variants (Bush et al. 1999) was analyzed. Eight pairs of sequences are identical, and only the 349 distinct sequences were used, each of 329 codons (987 nucleotides). This data set is referred to later as the large data set. A subset of the data containing 28 sequences (Yang et al. 2000), referred to later as the small data set, was analyzed as well. The HA gene encodes the major surface antigen, a target of neutralizing antibodies produced during infection or vaccination (Fitch et al. 1997). Previous studies suggest that the codon-based analysis is rather insensitive to the tree topology assumed (e.g., Yang et al. 2000). The present study thus does not examine the effect of tree topology. The tree topology for the small data set was obtained by ML (Yang et al. 2000), while that for the large data set was obtained by neighbor joining using the NG distances (Nei and Gojobori 1986; Saitou and Nei 1987).

The two data sets are analyzed under several models of variable ω ratios among sites, according to the recommendations of Yang et al. (2000). The site-homogeneous model M0 (one-ratio) assumes one ω for all sites. Model 1 (neutral) assumes a class of conserved sites with ω = 0 and another class of neutral sites with ω = 1. Model 2 (selection) adds a third class of sites with ω estimated. M3 (discrete) assumes a general discrete distribution, while the gamma model (M5) assumes a simple gamma distribution of ω over sites. Two other models used are M7 (beta), which assumes a beta distribution of ω, limited in the range (0, 1), and M8 (beta&ω), which adds an extra site class with ω estimated. The exact ML calculation and two approximate methods are used. The approximate methods calculate branch lengths either from pairwise estimates of d_S and d_N using the NG method (Nei and Gojobori 1986), or from a nucleotide-based analysis using the model of Yang (1996b). Apart from the way the branch lengths are obtained, there is no difference between the exact and approximate methods.

Table 1. Parameter estimates by approximate method I for the small data set

Model code	ℓ	Estimates of parameters	d_N/d_S	Positively selected sites
M0: one-ratio	−3140.38	$\omega = 0.392$	0.392	
M1: neutral	−3097.60	$p_0 = 0.657$ ($p_1 = 0.343$)	0.343	Not allowed
M2: selection	−3091.64	$p_0 = 0.653, p_1 = 0.337$ ($p_2 = 0.010$), $\omega_2 = 6.187$	0.400	135 226
M3: discrete	−3090.90	$p_0 = 0.760, p_1 = 0.234$ ($p_2 = 0.006$), $\omega_0 = 0.058, \omega_1 = 1.373, \omega_2 = 7.923$	0.416	Many
M7: beta	−3097.64	$p = 0.014, q = 0.027$	0.319	Not allowed
M8: beta&ω	−3091.63	$p_0 = 0.986, p = 0.013, q = 0.024$ ($p_1 = 0.014$), $\omega = 5.268$	0.386	135 226

The number of parameters in the ω distribution is 1, 1, 3, 5, 2, 4 for the five models, respectively. Estimates of the transition/transversion rate ratio κ are around 4.6. Positively selected sites include those with $P > 95\%$, with those with $P > 99\%$ in bold type

For the large data set, exact ML calculation involves heavy computation. The memory required ranges from 105MB (megabytes) for the site-homogeneous model (M0, one ratio, Goldman and Yang 1994) to 627MB for model M8 (beta&ω), which uses $K = 11$ site classes. The approximate methods, which fix branch lengths, require 105MB for all models. If ambiguity characters were removed, the space for conditional probabilities at the tips (53MB) can be saved for all models and methods. The computation was performed on Compaq AlphaStations. The calculation takes about 20 min for the site-homogeneous model (M0, one ratio) and 1–3 days for the complex site-class models such as M8 (beta&ω).

I consider the following aspects when comparing the exact and approximate methods: estimation of branch lengths, estimation of substitution parameters in the ω distribution, likelihood ratio test for the presence of sites under positive selection, and Bayes probabilities for site classes for identifying sites under selection.

Estimation of Branch Lengths

To examine how close the approximate estimates of branch lengths (y) are to the exact ML estimates (x), a linear regression is performed. As branch lengths under different models of variable ωs among sites are very similar (Yang et al. 2000), the exact estimates under model M3 (discrete) are used. If the approximation is perfect, the regression will be $y = x$ with $r^2 = 1$. For the small data set, the regressions are $y = 0.924x + 0.0013$ ($r^2 = 0.913$) for method I, and $y = 0.960x + 0.0001$ ($r^2 = 0.9995$) for method II. For the large data set, the regressions are $y = 0.8675x + 0.0005$ ($r^2 = 0.9029$) for method I, and $y = 0.9720x + 0.0001$ ($r^2 = 0.9849$) for method II. The approximations appear very good. In particular, approximate method II, which calculates branch lengths from a nucleotide-based analysis (Yang 1996b), gave very similar branch lengths to the exact ML calculation. It is also noted that the approximate branch lengths are underestimates and are much closer to the exact estimates under the site-homogeneous model (M0 one-ratio) than to those under M3. This pattern is expected as the two approximate methods tested here do not account for variable nonsynonymous rates among sites.

Estimation of Substitution Parameters and Likelihood Ratio Test of Positive Selection

ML estimates of parameters in the ω distribution for the small data set obtained by the two approximate methods are listed in Tables 1 and 2, respectively. They are very similar to estimates obtained by the exact ML method (not shown, but see Table 7 of Yang et al. 2000). Approximate method II, in particular, gave estimates essentially identical to the exact methods, with the log likelihood values <0.2 units worse for all models. Parameter estimates for the large data sets are listed in Table 3 for the exact method and in Tables 4 and 5 for the two approximate methods, respectively. Overall, the approximate methods produced estimates very similar to those of the exact method.

Table 6 lists the likelihood ratio statistics for two tests. The first compares the one-ratio model (M0) with the discrete model (M3). This is a test of the hypothesis that the ω ratio is identical among sites. The second test compares M7 (beta) against M8 (beta&ω), and directly tests for the presence of sites with $\omega > 1$. First, I note that the test statistics by the approximate and exact methods are similar for both the small and large data sets. Both tests are significant at the 1% level for the two data sets. As both M3 (discrete) and M8 (beta&ω) have classes with $\omega > 1$, the models provide significant evidence for positive selection, consistent with previous analyses (Fitch et al. 1997; Bush et al. 1999; Yang et al. 2000). Second, I note that the test statistics for the large data set are much greater than those for the small data set. This difference is clearly because a large number of sequences contain more information about ω ratios at individual sites and thus have greater power to detect positive selection.

Inference of Sites Under Selection

After branch lengths and substitution parameters are obtained, the Bayes theorem can be used to calculate the

Table 2. Parameter estimates by approximate method II for the small data set

Model code	ℓ	Estimates of parameters	d_N/d_S
M0: one-ratio	−3125.63	$\omega = 0.391$	0.391
M1: neutral	−3083.62	$p_0 = 0.662$ ($p_1 = 0.338$)	0.338
M2: selection	−3078.29	$p_0 = 0.657$, $p_1 = 0.333$ ($p_2 = 0.010$) $\omega_2 = 5.693$	0.391
M3: discrete	−3077.85	$p_0 = 0.746$, $p_1 = 0.247$ ($p_2 = 0.007$), $\omega_0 = 0.049$, $\omega_1 = 1.280$, $\omega_2 = 6.766$	0.400
M5: gamma	−3079.40	$\alpha = 0.234$, $\beta = 0.519$	0.399
M7: beta	−3083.65	$p = 0.014$, $q = 0.028$	0.317
M8: beta&ω	−3078.28	$p_0 = 0.987$, $p = 0.011$, $q = 0.021$ ($p_1 = 0.013$), $\omega = 5.069$	0.377

See notes for Table 1. Lists of positively selected sites are the same as in Table 1

Table 3. ML estimates of parameters for the large data set

Model	ℓ	Parameters	d_N/d_S	Positively selected sites
M0: one-ratio	−11,468.87	$\omega = 0.456$	$=\omega$	None
M1: neutral	−11,281.60	$p_0 = 0.439$ ($p_1 = 0.561$)	0.561	Not allowed
M2: selection	−11,123.59	$p_0 = 0.426$, $\omega_0 = 0$		**133 137 138 145 156**
		$p_1 = 0.516$, $\omega_1 = 1$		157 **159 186 193 194**
		$p_2 = 0.058$, $\omega_2 = 4.709$		**219 226** 246
M3: discrete	−11,009.63	$p_0 = 0.762$, $\omega_0 = 0.122$	0.459	Many
		$p_1 = 0.205$, $\omega_1 = 1.096$		
		$p_2 = 0.033$, $\omega_2 = 4.251$		
M7: beta	−11,129.58	$p = 0.249$, $q = 0.553$	0.310	Not allowed
m8: beta&ω	−11,016.94	$p_0 = 0.941$, $p = 0.377$, $q = 1.012$	0.441	133 **137 138 145 156**
		($p_1 = 0.059$), $\omega = 3.142$		157 **159 186 193 194**
				219 226

See notes for Table 1. Estimates of κ are around 3.6

Table 4. Parameter estimates by approximate method I for the large data set

Model	ℓ	Parameters	d_N/d_S	Positively selected sites
M0: one-ratio	−11,592.67	$\omega = 0.457$	$=\omega$	None
M1: neutral	−11,402.49	$p_0 = 0.451$ ($p_1 = 0.549$)	0.549	Not allowed
M2: selection	−11,243.04	$p_0 = 0.435$, $\omega_0 = 0$	0.760	**133 137 138 145 156**
		$p_1 = 0.512$, $\omega_1 = 1$		157 **159 186 193 194**
		$p_2 = 0.053$, $\omega_2 = 4.740$		**219 226** 246
M3: discrete	−11,129.45	$p_0 = 0.776$, $\omega_0 = 0.121$	0.423	Many
		$p_1 = 0.197$, $\omega_1 = 1.095$		
		$p_2 = 0.026$, $\omega_2 = 4.287$		
M5: gamma	−11,156.06	$\alpha = 0.284$, $\beta = 0.534$	0.479	**80 121 133 135 137 138 145 156 157 159 163** 172 **186 190 193 194 196** 197 **219 226 246 248** 275 **276** 310
M7: beta	−11,252.81	$p = 0.228$, $q = 0.570$	0.285	Not allowed
M8: beta&ω	−11,136.92	$p_0 = 0.951$, $p = 0.370$, $q = 1.020$	0.407	133 **137 138 145 156**
		($p_1 = 0.049$) $\omega = 3.174$		157 **159 186 193 194**
				219 226

See notes for Tables 1 and 3

posterior probabilities of site classes for each site (Nielsen and Yang 1998; Yang et al. 2000). Sites with high probabilities for classes with $\omega > 1$ are likely to be under positive selection. Sites under selection inferred this way are listed in Tables 1 and 2 for the two approximate methods for the small data set. Only two sites are identified by all models at the 95% level. There is no difference among the exact and approximate methods regarding this list. Sites inferred to be under positive selection in the large data set are listed in Tables 3–5. Again there is essentially no difference among the methods. The posterior means of ω for sites in the sequence

Table 5. Parameter estimates by approximate method II for the large data set

Model	ℓ	Parameters	d_N/d_S	Positively selected sites
M0: one-ratio	−11,470.76	$\omega = 0.457$	$=\omega$	None
M1: neutral	−11,283.33	$p_0 = 0.427$ ($p_1 = 0.573$)	0.573	Not allowed
M2: selection	−11,125.32	$p_0 = 0.422$, $\omega_0 = 0$	0.284	**133 137 138 145 156**
		$p_1 = 0.517$, $\omega_1 = 1$		157 **159 186 193 194**
		$p_2 = 0.061$, $\omega_2 = 4.708$		**219 226** 246
M3: discrete	−11,011.77	$p_0 = 0.762$, $\omega_0 = 0.121$	0.452	Many
		$p_1 = 0.205$, $\omega_1 = 1.090$		
		$p_2 = 0.032$, $\omega_2 = 4.241$		
M5: gamma	−11,033.65	$\alpha = 0.295$, $\beta = 0.532$	0.502	80 **121 133 135 137**
				138 145 156 157 159
				163 172 **186 190 193**
				194 196 219 226 246
				248 275 **276** 310
M7: beta	−11,132.03	$p = 0.239$, $q = 0.563$	0.298	Not allowed
M8: beta&ω	−11,018.91	$p_0 = 0.943$, $p = 0.375$, $q = 01.012$	0.434	133 **137 138 145 156**
		$p_1 = 0.057$, $\omega = 3.143$		157 **159 186 193 194**
				219 226

See notes for Tables 1 and 3

Table 6. Likelihood ratio statistics ($\Delta\ell$) for tests of positive selection

Method	M3 vs. M1 (d.f. = 5)	M8 vs. M7 (d.f. = 2)
Small data set		
Exact (from Yang et al. 2000)	47.88	5.43
Approximate I (Table 1)	49.48	6.01
Approximate II (Table 2)	47.78	5.37
Large data set		
Exact (Table 3)	459.24	112.64
Approximate I (Table 4)	463.22	115.89
Approximate II (Table 5)	458.99	113.12

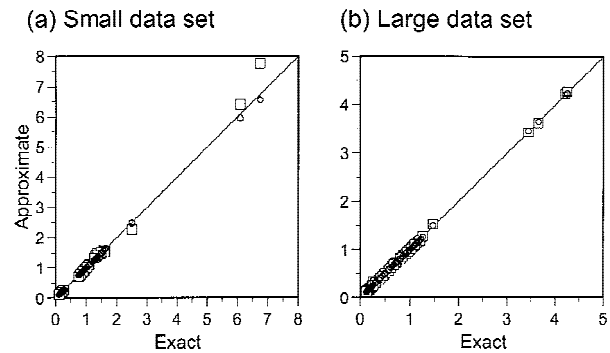

Fig. 2. Posterior means of ω at sites calculated using approximate methods I (□) and II (○), plotted against those calculated using the exact ML method. The discrete model (M3) is used.

are calculated under M3 and plotted in Fig. 2. Again we perform a linear regression of the approximate estimates (y) against the exact ones (x). For the small data set, the regressions are $y = 1.0701x - 0.0231$ ($r^2 = 0.9951$) and $y = 0.9847x + 0.0042$ ($r^2 = 0.9998$) for approximate methods I and II, respectively. For the large data set, they are $y = 1.0062x - 0.0018$ ($r^2 = 0.9999$) and $y = 0.9984x + 0.0006$ ($r^2 = 0.9999$). The correlations between the approximate and exact estimates of ω are much higher than the correlations between approximate and exact estimates of branch lengths, indicating that inference of sites under positive selection is somewhat robust to inaccuracies in branch length estimates. The posterior distribution and the posterior mean of ω for each site in the large data set calculated using the exact method are shown in Fig. 3.

Although constructed very differently, models M3 (discrete) and M8 (beta&ω) produced the same list of sites under positive selection for the large data set. We compare this list with previous analyses. Fitch et al. (1997) identified six sites under selection: 138, 145, 156, 186, 193, and 226 from an analysis of 254 sequences. Those sites are all inferred to be under positive selection at the 99% level by the likelihood analysis of this paper (Tables 3–5). In a later analysis of an extended data set including 357 sequences (the data set used in this paper), Bush et al. (1999) identified seven more sites under positive selection, of which two (133, 135) are in the lists of this paper while five (124, 142, 158, 190, 197) are not. Also four sites listed in Tables 3–5 (137, 157, 159, 219) are not in the list of Bush et al. (1999). Part of the differences may be due to the different treatment of the data, in particular, concerning counting of changes along tip branches on the tree. Overall the methods produce similar lists of sites under selection, although the significance values may be different.

Discussion

It is noted that the memory requirement of the exact ML calculation increases roughly linearly with the number of

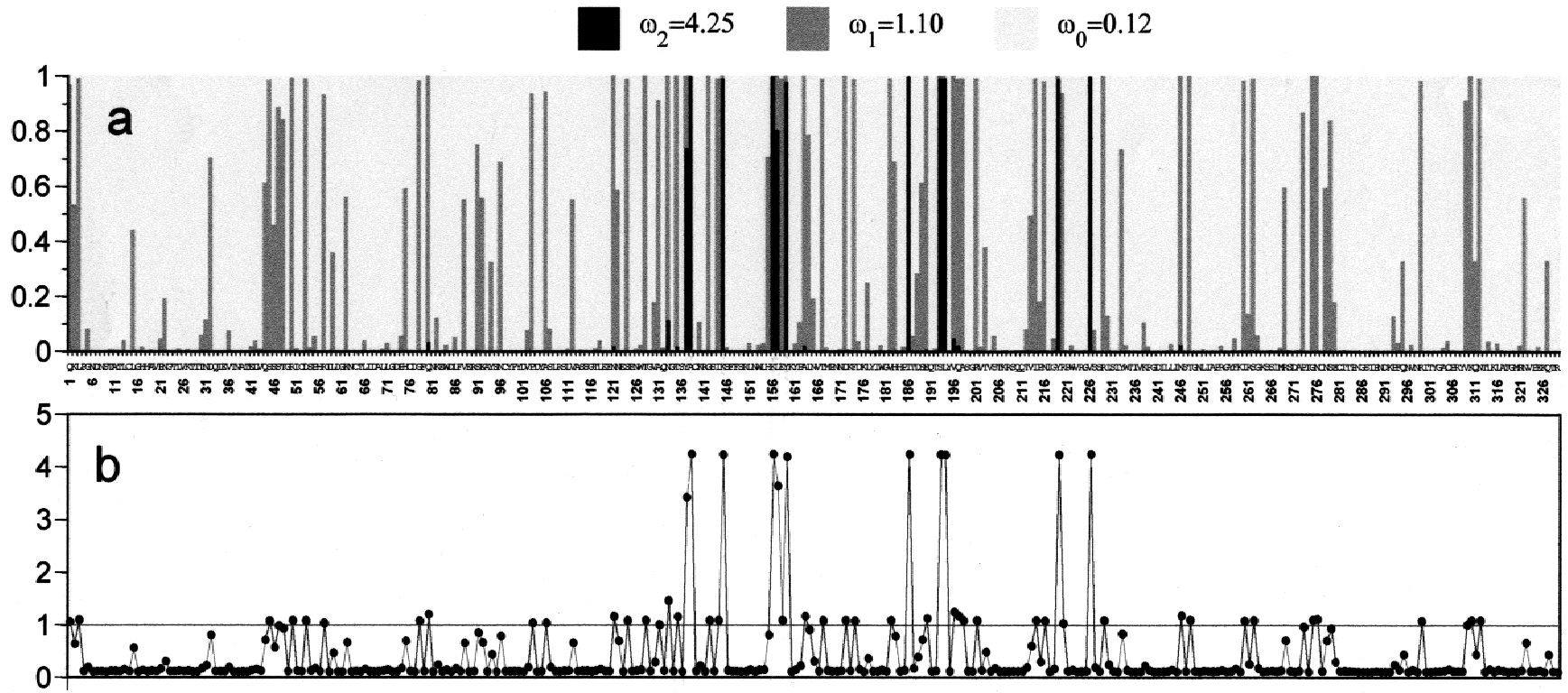

Fig. 3. Inference of sites under positive selection under M3 (discrete) for the large data set, using the Bayes theorem (Yang et al. 2000). Estimates of parameters under the model (Table 3) suggest three site classes in the proportions $p_0 = 0.762$, $p_1 = 0.205$, $p_2 = 0.033$ with $\omega_0 = 0.122$, $\omega_1 = 1.096$, $\omega_2 = 4.251$. Those proportions are the prior probabilities for each site, and the data at each site changes the prior into the posterior. The posterior probabilities of site classes **(a)** and the posterior mean of ω **(b)** for each site are plotted.

sequences in the data set (see Eq. 12). The amount of computation increases faster than linearly, but not by too much if the tree topology is fixed. With the improvement of computer power and algorithms, the exact method may soon be feasible for large data sets of hundreds of codon sequences. Nucleotide-based analysis requires much less memory and computation, and currently data sets of over a thousand sequences can be analyzed.

For data sets too large to handle by exact ML calculation, approximate methods for branch length estimation provides a useful alternative. Previous studies have employed approximate branch lengths obtained using least squares or parsimony methods for phylogeny reconstruction (e.g., Adachi and Hasegawa 1996). Errors in branch length estimates may have more impact on comparison of models than on comparison of tree topologies. Nevertheless, at least for the influenza data sets tested here, use of approximate branch lengths produced quite reliable results when compared with the exact ML method. A large amount of data may increase the power of the test so much that minor differences in likelihood are unlikely to change the conclusions. This is the case for the large data set (Table 6). The influenza virus gene sequences are quite similar. For more divergent sequences, it may be worthwhile to devise better algorithms for branch length estimation, for example, by taking into account variable nonsynonymous rates among sites.

Acknowledgments. I am grateful to Walter Fitch and Robin Bush for providing the sequence data analyzed in this paper. I thank Joe Felsenstein, Masami Hasegawa, and two anonymous referees for comments, and Joe Fletcher for running models M7 and M8 of Table 3 on a Compaq Alphaserver ES40. This study was supported by grants 31/G10434 and 31/MMI09806 from the Biotechnology and Biological Sciences Research Council (UK).

References

Adachi J, Hasegawa M (1996) MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. Comp Sci Monog 28:1–150. Institute of Statistical Mathematics, Tokyo

Bishop JG, Dean AM, Mitchell-Olds T (2000) Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. Proc Natl Acad Sci USA 97:5322–5327

Brent RP (1973) Algorithms for minimization without derivatives. Englewood Cliffs, New Jersey: Prentice-Hall

Bush RM, Fitch WM, Bender CA, et al. (1999) Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol Biol Evol 16:1457–1465

Crandall KA, Kelsey CR, Imamichi H, 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

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

Felsenstein J (1993) Phylip: Phylogenetic inference program, version 3. University of Washington, Seattle

Fitch WM, Bush RM, Bender CA, et al. (1997) Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci USA 94:7712–7718

Gill PE, Murray W, Wright MH (1981) Practical optimization. London: Academic Press

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

Grimmett GR, Stirzaker DR (1992) Probability and random processes. Oxford: Clarendon Press

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

Kishino H, Miyata T, Hasegawa M (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol 31:151–160

Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715–724

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426

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

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Suzuki Y, Gojobori T (1999) A method for detecting positive selection at single amino acid sites. Mol Biol Evol 16:1315–1328

Swofford DL (1999) PAUP*: phylogenetic analysis by parsimony* and other methods, version 4. Sanderland, MA: Sinauer Associates

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

Yang Z (1996a) Among-site rate variation and its impact on phylogenetic analyses. TREE 11:367–372

Yang Z (1996b) Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol 42:587–596

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Yang Z, Kumar S (1996) Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol Biol Evol 13:650–659

Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46:409–418

Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17:32–43

Yang Z, Nielsen R, Goldman N, et al. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449

Zanotto PM, Kallas EG, Souza RF, et al. (1999) Genealogical evidence for positive selection in the *nef* gene of HIV-1. Genetics 153:1077–1089