Codon-Substitution Models to Detect Adaptive Evolution that Account for Heterogeneous Selective Pressures Among Site Classes

Ziheng Yang and Willie J. Swanson†‡*

*Galton Laboratory, Department of Biology, University College London; †Department of Molecular Biology and Genetics, Cornell University; and ‡Department of Biology, University of California, Riverside

The nonsynonymous to synonymous substitution rate ratio ($\omega = d_N/d_S$) provides a sensitive measure of selective pressure at the protein level, with ω values <1, =1, and >1 indicating purifying selection, neutral evolution, and diversifying selection, respectively. Maximum likelihood models of codon substitution developed recently account for variable selective pressures among amino acid sites by employing a statistical distribution for the ω ratio among sites. Those models, called random-sites models, are suitable when we do not know a priori which sites are under what kind of selective pressure. Sometimes prior information (such as the tertiary structure of the protein) might be available to partition sites in the protein into different classes, which are expected to be under different selective pressures. It is then sensible to use such information in the model. In this paper, we implement maximum likelihood models for prepartitioned data sets, which account for the heterogeneity among site partitions by using different *ω* parameters for the partitions. The models, referred to as fixed-sites models, are also useful for combined analysis of multiple genes from the same set of species. We apply the models to data sets of the major histocompatibility complex (MHC) class I alleles from human populations and of the abalone sperm lysin genes. Structural information is used to partition sites in MHC into two classes: those in the antigen recognition site (ARS) and those outside. Positive selection is detected in the ARS by the fixed-sites models. Similarly, sites in lysin are classified into the buried and solvent-exposed classes according to the tertiary structure, and positive selection was detected at the solvent-exposed sites. The random-sites models identified a number of sites under positive selection in each data set, confirming and elaborating the results of the fixed-sites models. The analysis demonstrates the utility of the fixed-sites models, as well as the power of previous random-sites models, which do not use the prior information to partition sites.

Introduction

The nonsynonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$) provides a measure of selective pressure at the amino acid level. An ω ratio greater than 1 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 ω significantly exceeds 1. Previous studies have most often calculated synonymous (d_S) and nonsynonymous (d_N) rates by averaging over all codons (amino acids). As many amino acids in a functional protein may be under strong structural and functional constraints, the average d_N is rarely higher than the average d_S. As a result, this approach of averaging rates over the entire sequence has little power in detecting positive selection (e.g., Endo, Ikeo, and Gojobori 1996; Sharp 1997; Akashi 1999; Crandall et al. 1999).

Recently Nielsen and Yang (1998) and Yang et al. (2000) extended the model of codon substitution of Goldman and Yang (1994) (see also Muse and Gaut 1994) to account for variable selective pressures among sites in the sequence. A statistical distribution is assumed for ω ratios among sites. For example, the discrete model (M3) assumes three site classes, which have different ω ratios. The proportions and ω ratios for the site classes are estimated from the data by maximum likelihood. In such a model, we assume that there are several heterogeneous site classes but we do not know a priori which class each site is from. We refer to such models as random-sites models. Application of those models to real data sets has led to detection of positive selection in a number of genes, demonstrating the importance of accounting for variable selective pressures among sites (Zanotto et al. 1999; Bishop, Dean, and Mitchell-Olds 2000; Bielawski and Yang 2001; Fares et al. 2001; Ford 2001; Haydon et al. 2001; Peek et al. 2001; Swanson et al. 2001; see Yang and Bielawski 2000 for a review). Consistent with real data analysis, computer simulations also confirmed the power of those methods (Anisimova, Bielawski, and Yang 2001).

Sometimes prior information is available to partition sites into classes, which are expected to have different selective pressures and thus different ω ratios. In such cases, it is sensible to make use of such information and fit models that assign different ω ratios for site classes. For example, Hughes and Nei (1988) tested the hypothesis that amino acid residues at the antigen-recognition site (ARS) of the major histocompatibility complex (MHC) identified by Bjorkman et al. (1987*a*, 1987*b*) might be under diversifying selection. In this case, residues in the MHC can be partitioned into two classes: those in the ARS region and those outside, and two independent ω ratios can be used. Another possible use of such models is the combined analysis of multiple protein-coding genes from the same set of species to test for their similarities and differences in the substitution pattern. The models then have similarities to the relative-ratio test developed by Muse and Gaut (1997).

In this paper, we implement models that account for the heterogeneity of different site partitions, and re-

Key words: synonymous rate, nonsynonymous rate, positive selection, partitioned data, lysin, MHC, maximum likelihood, Bayes.

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(1):49–57. 2002

Table 1
Specifications of Models Implemented in This Paper

Model	Option G	Mgene	Parameters for Partitions (genes)	Number of Parameters
A	No	0	Same rates, κ, ω, and πs	$b + 2 + 9$
B	Yes	0	Different *r*s, same (κ, ω) and πs	$b + (g - 1) + 2 + 9$
C	Yes	2	Different *r*s, same (κ, ω), different πs	$b + (g - 1) + 2 + g \times 9$
D	Yes	3	Different *r*s and (κ, ω), same πs	$b + (g - 1) + g \times 2 + 9$
E	Yes	4	Different *r*s, (κ, ω), and πs	$b + (g - 1) + g \times 2 + g \times 9$
F	Yes	1	Separate analysis	$g \times (b + 2 + 9)$

NOTE.—The models are specified by the option variable G in the sequence data file and the variable Mgene in the control file in the PAML program package (Yang 1997). *b*: number of branches in the tree; *g*: number of genes or site partitions.

fer to them as the fixed-sites models. We apply the new models to two well-documented genes, the MHC class I gene (Hughes and Nei 1988, 1989; Hughes, Ota, and Nei 1990) and the abalone sperm lysin gene (Lee, Ota, and Vacquier 1995; Yang, Swanson, and Vacquier 2000).

Theory

As outlined by Yang (2001), implementation of the fixed-sites models requires simple modifications to the algorithm of Goldman and Yang (1994), which assumes that all sites in the sequence are under the same selective pressure and have the same ω ratio. The basic model of codon substitution specifies the relative substitution rate from codons *i* to *j* as

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

where κ is the transition/transversion rate ratio and π_j is the equilibrium frequency of codon *j*, calculated using the empirical nucleotide frequencies observed at the three codon positions, with nine parameters used (Goldman and Yang 1994).

When we apply the model to data of partitioned sites, we use different ω ratios, and thus different *Q* matrices, for sites from different partitions. Similarly we can allow other parameters to differ between site partitions. These models are structurally similar to models of nucleotide substitution of Yang (1996), which account for different transition/transversion rate ratios, different base frequencies, and different levels of among-site rate variation among prior partitions of sites, for example, the three codon positions. Here we also implement several models to accommodate different levels of site heterogeneity (table 1). The simplest model assumes that all sites in the sequence have the same substitution pattern with identical parameters (model A in table 1). Parameters in the model include the *b* branch lengths, the transition/transversion rate ratio κ, the nonsynonymous/synonymous rate ratio ω, and the nine parameters for the codon frequencies, with $b + 11$ parameters in total. The most complex model (model F in table 1) assumes that all site partitions have different substitution patterns with independent substitution parameters. This model is equivalent to analyzing data of different partitions as separate data sets and summing up the log-likelihood values. For *g* partitions, the model has $g \times (b + 11)$ parameters. Models B–E lie in between these two extremes, and assume proportional branch lengths among partitions. Branch lengths for partition *k* are r_k times those for the first partition ($r_1 = 1$). Thus $b + (g - 1)$, instead of $b \times g$, parameters are used to specify all branch lengths for the site partitions. Apart from the different substitution rates, model B (table 1) assumes homogeneity among partitions in the transition/transversion rate ratio κ, the nonsynonymous/synonymous rate ratio ω, and the codon frequencies. Model C assumes proportional branch lengths, identical κ and ω, but different codon frequencies among partitions. Model D assumes proportional branch lengths, different κ and ω, but identical codon frequencies among partitions. Model E assumes proportional branch lengths, different κ and ω, and different codon frequencies among partitions. These models are implemented in the PAML program package (Yang 1997); see table 1 for details.

The likelihood ratio test can be used to compare those models to test interesting hypotheses. For example, comparison between models A and B is a test of the hypothesis that the overall rate of nucleotide substitution is the same among partitions. The χ^2 distribution with d.f. $= g - 1$ can be used. Similarly, comparison of models C and E is a test of the hypothesis that κ and ω are identical across partitions. This comparison accounts for possible differences in codon usage among partitions.

Analysis of Class I MHC Alleles and Abalone Sperm Lysin Genes

We analyze two data sets to compare the fixed-sites models implemented in this paper and the random-sites models developed earlier (Nielsen and Yang 1998; Yang et al. 2000). The two data sets represent the most extensively characterized proteins that have been demonstrated to be under positive Darwinian selection: the class I major MHC locus (Hughes and Nei 1988) and the abalone sperm lysin (Lee, Ota, and Vacquier 1995; Yang, Swanson and Vacquier 2000). Multiple (≥25) sequences are available for both genes, permitting sensible phylogenetic comparisons, and crystal structures are available for representative proteins. Additionally, for the MHC, structural analyses have predicted sites that

Table 2
Log-likelihood Values and Parameter Estimates Under Random-sites Models for the Class I MHC Alleles

Model Code	p	ℓ	Estimates of Parameters	Positively Selected Sites
M0 (one-ratio).	392	−8225.16	$\hat{\omega} = 0.612$	None
M1 (neutral)	392	−7719.46	$\hat{p}_0 = 0.585$ ($\hat{p}_1 = 0.415$)	Not allowed
M2 (selection)	394	−7296.69	$\hat{p}_0 = 0.566$, $\hat{p}_1 = 0.332$, ($\hat{p}_2 = 0.102$), $\hat{\omega}_2 = 8.092$	**9F 24A 45M 62G 63E 67V 69A 70H 71S 77D 80T 81L 82R** *83G* **94T 95V 97R** *99Y* **113Y 114H 116Y 151H 152V 156L 163T 167W**
M3 (discrete).	396	−7226.51	$\hat{p}_0 = 0.780$, $\hat{p}_1 = 0.133$ ($\hat{p}_2 = 0.086$) $\hat{\omega}_0 = 0.069$, $\hat{\omega}_1 = 1.328$, $\hat{\omega}_2 = 6.048$	*9F 24A 45M 63E 67V 70H 71S 77D 80T 81L 82R 94T 95V 97R 99Y 113Y 114H 116Y 151H 152V 156L 163T 167W*
M7 (beta).	393	−7498.97	$\hat{p} = 0.103$, $\hat{q} = 0.354$	Not allowed
M8 (beta & ω)	395	−7232.68	$\hat{p}_0 = 0.900$, ($\hat{p}_1 = 0.100$), $\hat{p} = 0.168$, $\hat{q} = 0.710$, $\hat{\omega} = 5.122$	**9F 24A 45M** *62G* **63E 67V** *69A* **70H 71S 77D 80T 81L 82R 94T 95V 97R** *99Y* **113Y 114H 116Y 151H 152V 156L 163T 167W**

NOTE.—p is the number of parameters including $b = 381$ branch lengths. Estimates of κ range from 1.5 to 1.8. Sites inferred to be under positive selection at the 99% level are in bold and those at the 95% level are in italic. Sites listed for M3 are based on the third ω_2 class only. The reference sequence is from the structural file 1AKJ.

may be subjected to positive selection. These features allow for the amino acid sites in both proteins to be partitioned a priori, so that the models developed in this paper can be applied.

Class I MHC

The class I MHC glycoprotein recognizes and binds foreign peptides. The apparent selective force acting upon the MHC is to recognize and bind a large number of foreign peptides. Based on the crystal structure, different domains of the MHC have been characterized. The ARS is the cleft that binds foreign antigens (Bjorkman et al. 1987*a*, 1987*b*). The identification of the ARS enabled previous researchers to partition the data into ARS and non-ARS sites and to demonstrate positive selection in the ARS (Hughes and Nei 1988, 1989). Without partitioning the data, positive selection was not detected in pairwise comparisons averaging rates over the entire sequence. Therefore, the MHC makes an ideal test case for maximum likelihood analyses of partitioned data. We compiled and aligned 192 alleles of the human class I MHC from the A, B, and C loci. The alignment is available from the authors upon request. Alignment gaps were removed, with 270 codons left in each sequence. We used the maximum likelihood method to estimate pairwise distances under the codon-substitution model (Goldman and Yang 1994), and then used the neighbor-joining method (Saitou and Nei 1987) to construct a tree topology, which is used in later analysis. The tree topology was found to have little effect on the analysis in previous studies (e.g., Yang et al. 2000; Ford 2001), and in this paper we ignore the uncertainty of the tree topology.

First, we applied the random-sites models (Nielsen and Yang 1998; Yang et al. 2000) to the data. The results are presented in table 2. Model M0 assumes one ω ratio for all sites. The log likelihood is $\ell = -8225.16$, with the estimate $\hat{\omega} = 0.612$. This is an average over all sites in the protein and all lineages in the tree, and indicates the dominating role of purifying selection in the evolution of the MHC. Model M1 (neutral) assumes two site classes in the sequence: the conserved sites with $\omega_0 = 0$ and the neutral sites with $\omega_1 = 1$. This model has the same number of parameters as M0 (one-ratio) but fitted the data much better, with a log likelihood $\ell = -7719.46$. Model M2 (selection) adds another site class to M1 (neutral), with a free ω ratio estimated from the data, thus allowing for the possibility of positive selection. Parameter estimates suggest that about 10% of sites are under positive selection with $\hat{\omega}_2 = 8.1$ (table 2). This model fits the data much better than the neutral model; the test statistic is $2\Delta\ell = 2 \times (-7296.69 - (-7719.46)) = 845.54$, compared with the χ^2 distribution with d.f. = 2. Model M3 (discrete) assumes three site classes with the proportions (p_0, p_1, p_2) and ω ratios (ω_0, ω_1, ω_2) estimated from the data. The estimates suggest that the majority of sites are under purifying selection with $\hat{\omega}_0 = 0.07$, but about 9% of sites are under strong diversifying selection with $\hat{\omega}_2 = 6.0$. M3 fits the data significantly better than any of the simpler models M0, M1, or M2. Model M7 (beta) assumes a beta distribution of ω over sites. The beta distribution can take a variety of shapes although it is limited to the interval (0, 1). So it provides a flexible null model for testing positive selection. The estimated distribution $B(0.103, 0.354)$ has an extreme U shape, with most of the sites having ω close to either 0 or 1. Model M8 (beta & ω) adds an extra site class to M7 (beta) with a free ω ratio estimated from the data. The estimates suggest that about 10% of sites are under diversifying selection with $\hat{\omega} = 5.1$. The likelihood ratio test comparing M7 (beta) and M8 (beta & ω) has the statistic $2\Delta\ell = 2 \times (-7232.68 - [-7498.97]) = 2 \times 266.29 = 532.58$, much greater than a χ^2 significance value at d.f. = 2. Summing up, the random-sites models demonstrate extreme variability in selective pressure among sites in the MHC and the presence of a number of sites under diversifying selection. Sites inferred to be under positive

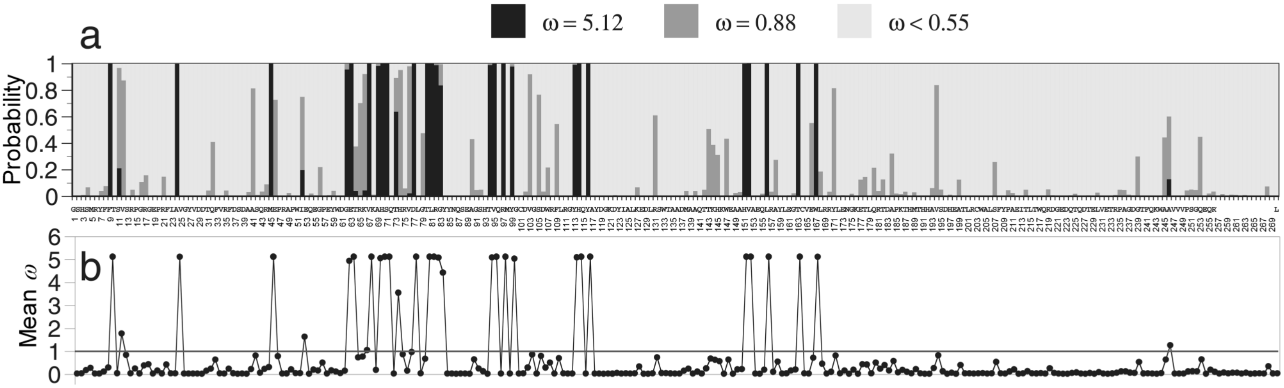

FIG. 1.—*a,* Posterior probabilities of site classes for sites along the MHC class I gene under the random-sites model M8 (beta & ω). Ten equal-probability categories are used to approximate the beta distribution (Yang et al. 2000), so that the model has 11 categories. The ω ratios are 0.00000, 0.00002, 0.00045, 0.00333, 0.01480, 0.04835, 0.12776, 0.28569, 0.54798, 0.88078, and 5.12163. Each of the first 10 categories has proportion 0.08998, where the last category has proportion 0.10019 (table 2). The first nine categories are collapsed into one category represented by $\omega < 0.55$. *b,* Posterior means of ω, calculated as the average of ω over the 11 site classes, weighted by the posterior probabilities. The amino acid sequence is from the structure file (Protein Data Bank file 1AKJ chain A; see fig. 2).

selection are listed in table 2. The posterior probabilities and posterior means for sites are shown in figure 1. Inferred sites are also mapped onto the crystal structure in figure 2. It is noteworthy that the sites inferred to be under positive selection are scattered along the primary sequence, but are all clustered in the ARS in the crystal structure (fig. 2).

To apply the fixed-sites models of this paper, we partitioned amino acid sites in the MHC into two classes: those located outside the ARS and those within, based on structural studies of Bjorkman et al. (1987*a*, 1987*b*) (see also Hughes and Nei 1988). The ARS class includes the following 57 sites: 5M, 7Y, 9F, 22F, 24A, 26G, 57P, 58E, 59Y, 61D, 62G, 63E, 64T, 65R, 66K, 67V, 68K, 69A, 70H, 71S, 72Q, 73T, 74H, 75R, 76V, 77D, 80T, 81L, 82R, 84Y, 95V, 97R, 99Y, 114H, 116Y, 143T, 145H, 146K, 147W, 149A, 150A, 151H, 152V, 154E, 155Q, 156L, 157R, 158A, 159Y, 161E, 162G, 163T, 165V, 166E, 167W, 169R, and 171Y. The site numbering is based on the sequence in the structure file 1AKJ (see fig. 1). The other 213 sites are located outside the ARS and lumped into the second site class.

Table 3 lists results obtained under the fixed-sites models. The simplest model (model A in table 3) assumes no site heterogeneity and gives $\ell_A = -8225.16$. Allowing for different substitution rates for the two partitions (model B in table 3) gave $\ell_B = -7790.10$. This is a dramatic improvement of 435.06 log-likelihood units upon adding a single parameter (r_2). The estimate r_2 indicates that the substitution rate in the ARS is 6.5 times as high as outside the ARS. Model C further allows for different codon frequencies for the two partitions, by using nine additional parameters for base frequencies at the three codon positions. The log likelihood increased by $\ell_C - \ell_B = -7767.77 - (-7790.10) = 22.33$. While statistically significant, this is not a very big improvement. Model D uses different κ and ω but the same codon frequencies for the two partitions. It has two more parameters than model B and fits the data much better; the likelihood ratio statistic is $2\Delta\ell = 2(\ell_D - \ell_B) = 2 \times ([-7691.57] - [-7790.10]) = 197.06$. Variation in κ and ω between the partitions is much more important to the fit of the model than variation in the codon frequencies. Model E assumes different κ and ω as well as different codon frequencies for the two partitions, and fits the data significantly better than any of the simpler models. Parameter estimates under model E are similar to those under model D. They all suggest that the ω ratio is very different in the two partitions. The non-ARS sites are under purifying selection with $\hat{\omega}_1 = 0.23$, whereas the ARS sites are under diversifying selection with $\hat{\omega}_2 = 1.9$. Like the comparison between models B and D, comparison between models C and E leads to rejection of model C, with $2\Delta\ell = 2(\ell_E - \ell_C) = 191.70$, indicating that κ and ω are different between the partitions. Model F is the separate analysis. Despite its use of 381 × 2 branch lengths for the two partitions, many of which are zero, the model fits the data significantly better than models for combined analysis which assume proportional branch lengths (models B, C, D, and E). For example, the test statistic for comparing models E and F is $2\Delta\ell = 492.34$, and $P < 0.0001$ with d.f. = 380. Nevertheless, estimates of parameters such as κ and ω are highly similar to those obtained in the combined analyses. The tree length, i.e., the sum of

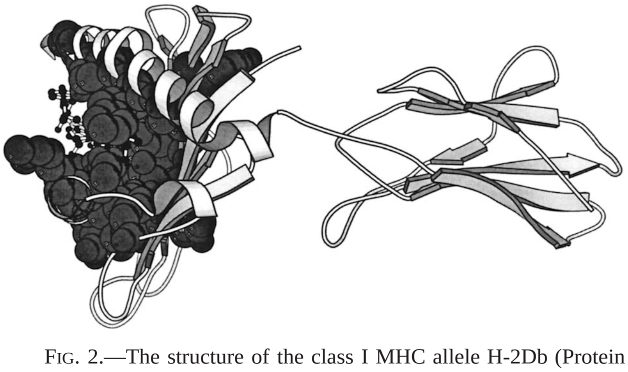

FIG. 2.—The structure of the class I MHC allele H-2Db (Protein Data Bank file 1AKJ chain A; Gao et al. 1997), with a bound antigen shown in stick and ball format. Amino acid residues identified to be under positive selection by the random-site model M8 (beta & ω) are shown in spacefill, and all fall in the ARS domain.

Table 3
Log-Likelihood Values and Parameter Estimates Under Fixed-Sites Models for the Class I MHC Alleles

Model	p	ℓ	$\hat{r}_2$	$\hat{\kappa}$	$\hat{\omega}$
A (homogeneous model)	392	−8225.16	1	1.76	0.612
B (different rs)	393	−7790.10	6.54	1.73	0.583
C (different rs and πs)	402	−7767.77	6.34	1.68	0.575
D (different rs, κ and ω)	395	−7691.57	6.35	$\hat{\kappa}_1 = 2.48$ $\hat{\kappa}_2 = 1.37$	$\hat{\omega}_1 = 0.228$ $\hat{\omega}_2 = 1.938$
E (different rs, κ and ω, and πs) . . .	404	−7671.92	6.25	$\hat{\kappa}_1 = 2.54$ $\hat{\kappa}_2 = 1.34$	$\hat{\omega}_1 = 0.232$ $\hat{\omega}_2 = 1.860$
F (separate analysis)	784	−7425.75	6.18	$\hat{\kappa}_1 = 2.57$ $\hat{\kappa}_2 = 1.33$	$\hat{\omega}_1 = 0.232$ $\hat{\omega}_2 = 1.853$

NOTE.—p: Number of parameters including $b = 381$ branch lengths. The two partitions are the non-ARS sites and the ARS sites. r_2 is the rate of the second site partition relative to the rate of the first partition ($r_1 = 1$).

branch lengths along the tree, for the first partition (sites outside the ARS) is 1.957 nucleotide substitutions per codon, or $\hat{d}_S = 1.789$ synonymous substitutions per synonymous site and $\hat{d}_N = 0.414$ nonsynonymous substitutions per nonsynonymous site. At the ARS, the tree length is 12.087 nucleotide substitutions per codon, or $\hat{d}_S = 2.317$ and $\hat{d}_N = 4.297$. Therefore, the synonymous rates are similar between the two partitions, and the over sixfold difference in substitution rate between the two partitions is mainly caused by the accelerated nonsynonymous rate at the ARS.

To test whether the ω ratio at the ARS is significantly different from 1, we recalculated the log-likelihood values in models D, E, and F by fixing $\omega_2 = 1$. If the ARS sites only are analyzed under the one-ratio model (model F; Goldman and Yang 1994), the log likelihood is −3857.64 when ω is a free parameter and −3866.58 when $\omega = 1$ is fixed. Thus the likelihood ratio test statistic is $2\Delta\ell = 2 \times ([-3857.64] - [-3866.58]) = 17.88$, with $P = 2.4 \times 10^{-5}$ at d.f. = 1 (table 4, model F). Models D and E analyze the two partitions as one combined data set. When $\omega_2 = 1$ is fixed in model D, the log likelihood is $\ell = -7702.55$, so the test statistic $2\Delta\ell = 2 \times ([-7691.57] - [-7702.55]) = 21.95$. Under model E, the test statistic $2\Delta\ell = 2 \times ([-7671.92] - [-7681.25]) = 18.66$. All these tests, which make different assumptions about differences between the two site partitions, reject the null hypothesis and suggest that the ratio ω_2 at the ARS is significantly greater than 1 (table 4).

The fixed-sites and random-sites models are not nested and cannot be compared using a simple χ^2 distribution. Nevertheless, the log-likelihood values are comparable between the two classes of models. Note that in a fixed-sites model, the probability of observing data at a site is calculated using the ω ratio for the partition the site is from. In a random-sites model, the probability is calculated as an average over all site classes (Nielsen and Yang 1998; Yang et al. 2000). Because averaging over unlikely site classes reduces the probability, we expect the fixed-sites models to have much higher likelihood values than the random-sites models. However, results of tables 2 and 3 suggest the opposite, and the fixed-sites models fit the data much more poorly than the random-sites models. For example, the fixed-sites model E in table 3 has 404 parameters and $\ell_E = -7671.92$, whereas the random-sites model M8 (beta & ω) in table 2 has 395 parameters but a much higher log-likelihood value, $\ell = -7232.68$, with a difference of 439.24.

The poorer performance of the fixed-sites models appears to be mainly caused by inclusion of conserved sites in the list of the 57 ARS sites. We note that structural studies permit the identification of sites potentially involved in antigen binding, but do not expect all of them to be under diversifying selection in the data set examined. The random-sites model M8 (beta & ω) identified 25 sites to be under positive selection (table 2), out of which 22 are in the list of ARS sites. The three sites that are not in the list are 45M, 94T, and 113Y. These sites are located in the ARS domain, although not in the binding cleft, and might also be involved in specificity of binding foreign peptides. Previous studies demonstrated that antibody specificity can be mediated by both variable loops and substitutions on the protein framework that do not have direct contact with the antigen (Foote and Winter 1992). The results here suggest a similar process may be occurring at these sites in the MHC. There are 35 sites in the ARS partition that are not identified to be under positive selection by the random-sites models. Of them, site 73T has posterior probability $P = 0.64$ and posterior mean $\hat{\omega} = 3.6$, and is quite likely to be under positive selection (fig. 1). Sites 64T, 66K, 74H, 75R, 76V, and 171Y all have posterior means $\hat{\omega} > 0.8$ and are possibly under positive selection but not detected by the random-sites models because of lack of information in the data at these sites. Sites 5M, 22F, 26G, 57P, 72Q, 84Y, 146K, 154E, 159Y, 165V, and 169R have posterior probabilities close to zero and pos-

Table 4
Likelihood Ratio Test Statistic ($2\Delta\ell$) for Testing the Null Hypothesis that $\omega_2 = 1$ for the ARS Sites in the Class 1 MHC

Model Assumed	$2\Delta\ell$
D: proportional branch lengths, same πs	21.95*
E: proportional branch lengths, different πs	11.20*
F: using ARS sites alone	17.88*

* $P < 0.01$, d.f. = 1.

Table 5
Log-likelihood Values and Parameter Estimates Under Two Random-Sites Models for the Abalone Sperm Lysin Gene

Model	p	ℓ	Estimates of Parameters	Positively Selected Sites
M7: beta	59	−4,472.16	$\hat{p} = 0.281$, $\hat{q} = 0.210$	Not allowed
M8: beta & ω.	61	−4,410.57	$\hat{p}_0 = 0.733$, $\hat{p} = 0.305$, $\hat{q} = 0.281$, $\hat{p}_1 = 0.267$, $\hat{\omega} = 2.990$	**4H 7E 9K 10F 12N** *14A* **32V 33K 36R** *41T* **44T** *64N* **67L 70N 74D 83G** *86T* **87R 113K 120E** *126P 127A* **132K**

NOTE.—p: Number of parameters including $b = 47$ branch lengths in the tree. Estimates of κ are 1.3 under M7 and 1.6 under M8. Sites inferred to be under positive selection at the 99% level are listed in bold and those at the 95% level are in italic. The reference sequence is that of the red abalone.

terior mean $\omega < 0.1$ (fig. 1). These sites are most likely to be under strong purifying selection. Indeed, sites 57P, 72Q, 154E, 165V, and 169R point away from the antigen binding cleft and were predicted not to be involved in direct antigen binding in the original MHC structural analysis (Bjorkman et al. 1987*b*).

Overall, these comparisons demonstrate the consistency of the fixed-sites and random-sites models and, in particular, the utility of the random-sites models even when structural information is available. They also highlight the power of predicting binding sites by incorporating both structural and evolutionary information.

It is also interesting to compare the results of table 2 (see also fig. 2) with those of Swanson et al. (2001), who applied the random-sites models to a dataset of only six MHC alleles. The smaller data set included the signal sequence and additional C-terminal sequence, which were removed in this paper because these regions were not sequenced in all 192 alleles analyzed. Under the numbering system of this paper, this analysis identified 12 sites at the 50% level: 45M, 62G, 63E, 66K, 67V, 70H, 71S, 97R, 114H, 116Y, 151H, and 156L. All but one site (site 66K) are in the list of this paper (table 2). It is remarkable that all sites identified in both studies are clustered in the ARS domain. At the 95% level, only two sites (114H and 156L) were identified in the small data set, compared with 25 sites in this paper. This comparison demonstrates the dramatic improvement in the power of the method with the increase of the number of sequences used, consistent with the simulation study of Anisimova, Bielawski, and Yang (2001). We suggest that more sites might be under positive selection in the MHC than identified in this paper.

Abalone Sperm Lysin

Abalones are large marine gastropod mollusks that exhibit external fertilization, with sperm and eggs released directly into seawater where fertilization occurs. Despite many of the species having overlapping breeding seasons and habitats, the species remain distinct. One barrier to cross-species fertilization is the species-specific interaction of sperm and eggs, which can be quantitatively demonstrated in the laboratory (e.g., Lyon and Vacquier 1999). The molecules involved in the species-specific interaction have been characterized extensively (reviewed in Vacquier et al. 1999). Abalone sperm lysin is a 16-kDa protein localized in the sperm acrosome granule. Upon exocytosis, lysin dissolves a hole in the egg vitelline envelope (VE) in a nonenzymatic and species-specific manner. Lysin binds to and unravels the fibrous VE by disrupting hydrogen bonds and hydrophobic interactions of its receptor VERL (Swanson and Vacquier 1997, 1998). The crystal structures of the red (*Haliotis rufescens*) and green (*H. fulgens*) abalone have been determined (Shaw et al. 1995; Kresge, Vacquier, and Stout 2000*a*, 2000*b*). The sperm lysin genes of 25 abalone species were sequenced and analyzed by Lee, Ota, and Vacquier (1995), and strong diversifying selection was demonstrated at a number of amino acid sites in lysin, particularly in closely related sympatric species (Yang, Swanson, and Vacquier 2000). The sequence data used in this paper are the same as those analyzed by Lee, Ota, and Vacquier (1995) and Yang et al. (2000), except that an alignment gap between residues 133 and 134 in the original alignment is deleted in this paper, so that 134 codons are in each sequence. We use the phylogeny estimated by Lee, Ota, and Vacquier (1995).

Extensive analysis of the data under random-sites models was performed by Yang et al. (2000). In this paper, we present results obtained under models M7 (beta) and M8 (beta & ω) only (table 5). Parameter estimates are essentially identical to those in Yang et al. (2000), but the log-likelihood values are quite different, because of the removed site. Estimates under model M8 (beta & ω) suggest that many sites are highly conserved, but as many as 27% of sites are under diversifying selection with $\hat{\omega}_2 = 3.0$. The likelihood ratio test comparing these two models suggests that the difference is statistically significant; the test statistic is $2\Delta\ell = 2(\ell_1 - \ell_0) = 2 \times ([-4410.57] - [-4472.16]) = 123.18$, compared with the χ^2 distribution with d.f. = 2. Sites inferred to be under positive selection are listed in table 5. The lysin structure of the red abalone (*H. rufescens*), with sites identified to be under positive selection mapped onto it, was presented in Yang, Swanson, and Vacquier (2000).

As lysin is a surface-active molecule, we hypothesize that solvent-exposed residues in lysin might be subjected to positive selection, whereas the buried residues would be conserved in order to maintain the protein structure. To test this hypothesis, we partitioned the 134 sites in lysin into two classes, the buried sites and the solvent-exposed sites. Solvent accessibility is calculated from the red abalone lysin structure (1LIS in Protein Data Bank) using the program GETAREA (http://www.scsb.utmb.edu/cgi-bin/get_a_form.tcl; Fraczkiewicz and Braun 1998). The first partition includes the following 46 bur-

Table 6
Log-likelihood Values and Parameter Estimates for the Abalone Sperm Lysin Gene Under the Fixed-sites Models

Model	p	ℓ	$\hat{r}_2$	$\hat{\kappa}$	$\hat{\omega}$
A (homogeneous)	58	−4,627.03	1	1.58	0.955
B (different *r*s)	59	−4,549.99	2.76	1.55	0.947
C (different *r*s and πs)	68	−4,490.07	2.61	1.60	0.997
D (different *r*s, κ and ω)	61	−4,532.12	2.69	$\hat{\kappa}_1 = 1.73$ $\hat{\kappa}_2 = 1.51$	$\hat{\omega}_1 = 0.393$ $\hat{\omega}_2 = 1.248$
E (different *r*s, κ and ω, and πs)	70	−4,473.88	2.54	$\hat{\kappa}_1 = 1.93$ $\hat{\kappa}_2 = 1.51$	$\hat{\omega}_1 = 0.452$ $\hat{\omega}_2 = 1.277$
F (separate analysis)	116	−4,454.85	2.52	$\hat{\kappa}_1 = 1.96$ $\hat{\kappa}_2$ 1.51	$\hat{\omega}_1 = 0.452$ $\hat{\omega}_2 = 1.268$

NOTE.—p: Number of parameters including $b = 47$ branch lengths in the tree. The two partitions are for the buried and solvent exposed sites. r_2 is the rate of the second site partition relative to the rate of the first partition ($r_1 = 1$).

ied residues: 11L, 15F, 16E, 18A, 19L, 20K, 23I, 24I, 26G, 27F, 31L, 34W, 35L, 39G, 42L, 49A, 50L, 52F, 53V, 54N, 55R, 58M, 59Q, 62W, 65Y, 66M, 69I, 73I, 84D, 85Y, 88L, 89G, 92I, 93G, 98M, 102Y, 105L, 111I, 112P, 114Y, 118M, 121I, 122N, 129V, 131V, and 133Y. The remaining 88 residues are in the solvent-exposed class.

The results obtained under the fixed-sites models are shown in table 6. Model A, which assumes the same parameters in the two partitions, gave $\ell_A = -4627.03$. Model B allows the overall rates to differ and fits the data much better than model A; the likelihood ratio test statistic is $2\Delta\ell = 2 \times ([-4549.99] - [-4627.03]) = 154.08$, compared with the χ^2 distribution with d.f. = 1. The rate at the solvent-exposed sites is 2.8 times as high as at the buried sites (r_1:$\hat{r}_2$ = 1:2.755). Model C allows further for different codon frequencies for the two partitions, determined by the nucleotide frequencies at the three codon positions. This model fits the data much better than model B ($2\Delta\ell = 119.84$, d.f. = 9), suggesting that the codon usage patterns are indeed different at the buried and exposed sites. Model D assumes the same codon frequencies but different transition/transversion rate ratio κ and nonsynonymous/synonymous rate ratio ω. This model fits the data better than model B ($2\Delta\ell = 35.74$, d.f. = 2). The estimates are $\hat{\kappa}_1 = 1.7$ and $\hat{\omega}_1 = 0.39$ for the buried sites and $\hat{\kappa}_2 = 1.5$ and $\hat{\omega}_2 = 1.25$ for the solvent-exposed sites (table 6). Whereas estimates of κ are similar between the partitions, estimates of ω are very different. As hypothesized, buried sites are under strong purifying selection, and solvent-exposed sites appear to be under diversifying selection. Unlike the MHC data set, allowing for different codon frequencies (model C) improves the fit of the model more than allowing for different κ and ω (model D). This pattern might be the result of different amino acid compositions at the buried and exposed sites. Model E allows different κ and ω as well as different codon frequencies between partitions, and fits the data better than any of the simpler models. The model gave similar estimates of parameters as model D (table 6). Model F is equivalent to separate analysis of the two partitions. It is not significantly better than model E; the statistic is $2\Delta\ell = 38.06$, and $P = 0.79$, with d.f. = 46. So it is acceptable to use 47 + 1 instead of 47 × 2 parameters for branch lengths in the two partitions. Parameter estimates under model F are similar to those obtained in the combined analyses (models B–E). The tree length for the buried sites is 3.96 nucleotide substitutions per codon, or $\hat{d}_S = 2.20$ synonymous substitutions per synonymous site and $\hat{d}_N = 0.99$ nonsynonymous substitutions per nonsynonymous site. The tree length for the solvent-exposed sites is 9.98, or $\hat{d}_S = 2.76$ and $\hat{d}_N = 3.50$. Thus the 2.5 times rate difference between the two partitions is mainly caused by the accelerated nonsynonymous rate at the exposed sites.

To test whether the ω ratio at the solvent-exposed sites is significantly greater than 1, we recalculated the log-likelihood values in models D, E, and F by fixing $\omega_2 = 1$. In an analysis of the exposed sites only (model F), the log likelihood is −3517.01 when ω is estimated as a free parameter and −3519.62 when ω = 1 is fixed. Thus the likelihood ratio statistic for testing the null hypothesis $\omega_2 = 1$ is $2\Delta\ell = 5.23$, with $P = 0.022$ at d.f. = 1 (table 7). Models D and E analyze the two partitions as one combined data set and have the test statistics to be 4.53 and 5.57, respectively (table 7). So, whatever our assumptions about possible differences between the two partitions, we reject the hypothesis $\omega_2 = 1$ at $1\% < P < 5\%$, and conclude that the solvent-exposed sites in lysin are under diversifying selection with $\omega_2 > 1$ (table 7).

All sites predicted by the random-sites models to be under positive selection are located on the surface of lysin and, therefore, included in the solvent exposed class. Similar to the analysis of the MHC data set, the fixed-sites models fit the data more poorly than the random-sites models, judged by their log-likelihood values. The main reason for this difference appears to be that

Table 7
Likelihood Ratio Test Statistic ($2\Delta\ell$) and P Value for Testing the Null Hypothesis that $\omega_2 = 1$ at the Solvent Exposed Sites in Lysin

Model Assumed	$2\Delta\ell$	P
D: proportional branch lengths, same πs	5.57	0.018
E: proportional branch lengths, different πs	4.53	0.033
F: using the solvent exposed sites only	5.23	0.022

some exposed sites are under purifying rather than positive selection.

Discussions

Comparison of Fixed-Sites and Random-Sites Models

The analyses of both the MHC and the lysin data sets demonstrate the utility of the new fixed-sites models implemented in this paper. In both genes, the ω ratio averaged over all sites in the sequence is less than 1. However, positive selection is detected when structural information is used to identify sites that might be expected to be under positive selection, and an independent ω ratio is assigned to the partition of such sites in the likelihood model. Perhaps more remarkable is the power of the random-sites models, which do not use structural information to partition sites. In both genes, the random-sites models provided even better fit to the data than the fixed-sites models, indicated by their higher log-likelihood values. This discrepancy appears to be caused by the inclusion of conserved sites in the site partition expected to be under positive selection used in the fixed-sites models, so that there is still substantial variation in selective pressure among sites within the same partition. In terms of statistical significance for detecting positive selection, we suspect that the fixed-sites models will seldom be more powerful than the random-sites models. To obtain significant results about positive selection by the fixed-sites models, it will be necessary to have reliable information to partition sites and a number of sites in the partition under fairly strong positive selection. In such cases, the random-sites models are unlikely to fail.

We note that in the MHC data set, 22 of the 25 sites identified by the random-sites models to be under positive selection are in the list of sites in the ARS, whereas the other three sites are in the ARS domain. In the lysin data set, all sites identified by the random-sites models are in the partition of exposed sites. Such consistency between the two classes of models validates the biological hypothesis used to partition sites a priori and also the reliability of the random-sites models. We suggest that the random-sites models are useful whether or not prior information is available to partition sites in the sequence. However, it should be emphasized that identification of sites under positive selection using the Bayes theorem requires simultaneous inferences at all sites in the sequence. Whereas the accuracy at one site might be high as indicated by the posterior probability, it is very unlikely for all sites to be identified correctly. Furthermore, the empirical Bayes procedure we used does not account for the sampling errors in parameter estimates, and the posterior probability calculations might be sensitive to parameters in the ω distribution (Yang and Bielawski 2000). Those problems may be serious when the analyzed data set is small and contains only a few highly similar sequences, with little information to estimate parameters in the ω distribution. Thus we suggest that caution be exercised and the inferred sites be considered hypotheses to be verified by experimental investigation.

Analysis of Data from Multiple Genes

We envisage that one major use of the fixed-sites models is to test for similarities and differences in the evolutionary process among different genes. When sequences from multiple protein-coding genes are available for the same set of species, they can be analyzed as a combined data set, with their differences in the substitution pattern accounted for. Interesting hypotheses concerning differences among genes in the selective pressure indicated by the ω ratio can then be tested. In this regard, some variations to the models we implemented here might be more interesting. For example, one such model might have a homogeneous synonymous substitution rate and variable nonsynonymous rates among genes. Another model might assume proportional branch lengths at the synonymous site and freely variable branch lengths at the nonsynonymous site among the genes. It might also be worthwhile to decouple κ and ω. In this paper, these two parameters are either both homogeneous or both different among genes. Analyses of this paper did not assume a molecular clock, so that the overall rate varies among branches. Models that enforce the molecular clock at the synonymous site but do not enforce the clock at the nonsynonymous site might be interesting. We note that some similar models have been developed by Muse and Gaut (1997) in their pioneering work, and further implementation of such likelihood models is straightforward.

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. **18**:1585–1592.

Bielawski, J. P., and Z. Yang. 2001. Positive and negative selection in the DAZ gene family. Mol. Biol. Evol. **18**:523–529.

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.

Bjorkman, P. J., M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger, and D. C. Wiley. 1987*a*. The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. Nature **329**:512–518.

Bjorkman, P. J., M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger, and D. C. Wiley. 1987*b*. Structure of the class I histocompatibility antigen, HLA-A2. Nature **329**: 506–512.

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.

FOOTE, J., and G. WINTER. 1992. Antibody framework residues affecting the conformation of the hypervariable loops. J. Mol. Biol. **224**:487–499.

FORD, M. J. 2001. Molecular evolution of transferrin: evidence for positive selection in salmonids. Mol. Biol. Evol. **18**: 639–647.

FRACZKIEWICZ, R., and W. BRAUN. 1998. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. J. Comp. Chem. **19**: 319–333.

GAO, G. F., J. TORMO, U. C. GERTH, J. R. WYER, A. J. MCMICHAEL, D. I. STUART, J. I. BELL, E. Y. JONES, and B. K. JAKOBSEN. 1997. Crystal structure of the complex between human CD8alpha(alpha) and HLA-A2. Nature **387**: 630–634.

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.

HUGHES, A. L., and M. NEI. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**:167–170.

HUGHES, A. L., and M. NEI. 1989. Evolution of the major histocompatibility complex: independent origin of nonclassical class I genes in different groups of mammals. Mol. Biol. Evol. **6**:559–579.

HUGHES, A. L., T. OTA, and M. NEI. 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Mol. Biol. Evol. **7**:515–524.

KRESGE, N., V. D. VACQUIER, and C. D. STOUT. 2000*a*. 1.35 and 2.07 Å resolution structures of the red abalone sperm lysin monomer and dimer reveal features involved in receptor binding. Acta Crystallogr. **56**:34–41.

———. 2000*b*. The high resolution crystal structure of green abalone sperm lysin: implications for species-specific binding of the egg receptor. J. Mol. Biol. **296**:1225–1234.

LEE, Y.-H., T. OTA, and V. D. VACQUIER. 1995. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Mol. Biol. Evol. **12**:231–238.

LYON, J. D., and V. D. VACQUIER. 1999. Interspecies chimeric sperm lysins identify regions mediating species-specific recognition of the abalone egg vitelline envelope. Dev. Biol. **214**:151–159.

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.

MUSE, S. V., and B. S. GAUT. 1997. Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test. Genetics **146**:393–399.

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.

PEEK, A. S., V. SOUZA, L. E. EGUIARTE, and B. S. GAUT. 2001. The interaction of protein structure, selection, and recombination on the evolution of the type-1 fimbrial major subunit (*fimA*) from *Escherichia coli*. J. Mol. Evol. **52**:193–204.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

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

SHAW, A., P. A. FORTES, C. D. STOUT, and V. D. VACQUIER. 1995. Crystal structure and subunit dynamics of the abalone sperm lysin dimer: egg envelopes dissociate dimers, the monomer is the active species. J. Cell Biol. **130**:1117–1125.

SWANSON, W. J., and V. D. VACQUIER. 1997. The abalone egg vitelline envelope receptor for sperm lysin is a giant multivalent molecule. Proc. Natl. Acad. Sci. USA **94**:6724–6729.

SWANSON, W. J., and V. D. VACQUIER. 1998. Concerted evolution in an egg receptor for a rapidly evolving abalone sperm protein. Science **281**:710–712.

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.

VACQUIER, V. D., W. J. SWANSON, E. C. METZ, and C. D. STOUT. 1999. Acrosomal proteins of abalone spermatozoa. Adv. Dev. Biochem. **5**:49–81.

YANG, Z. 1996. 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. 2001. Adaptive molecular evolution. Pp. 327–350 *in* D. BALDING, M. BISHOP, and C. CANNINGS, eds. Handbook of statistical genetics. Wiley, New York.

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

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.

YANG, Z., W. J. SWANSON, and V. D. VACQUIER. 2000. Maximum likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Mol. Biol. Evol. **17**:1446–1455.

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.

BRANDON GAUT, reviewing editor

Accepted August 27, 2001