DOI: 10.1534/genetics.102.010959

Detecting Selection in Noncoding Regions of Nucleotide Sequences

Wendy S. W. Wong[1] and Rasmus Nielsen

Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14850

Manuscript received October 30, 2002
Accepted for publication December 31, 2003

ABSTRACT

We present a maximum-likelihood method for examining the selection pressure and detecting positive selection in noncoding regions using multiple aligned DNA sequences. The rate of substitution in noncoding regions relative to the rate of synonymous substitution in coding regions is modeled by a parameter ζ. When a site in a noncoding region is evolving neutrally $\zeta = 1$, while $\zeta > 1$ indicates the action of positive selection, and $\zeta < 1$ suggests negative selection. Using a combined model for the evolution of noncoding and coding regions, we develop two likelihood-ratio tests for the detection of selection in noncoding regions. Data analysis of both simulated and real viral data is presented. Using the new method we show that positive selection in viruses is acting primarily in protein-coding regions and is rare or absent in noncoding regions.

MUCH attention has recently been given to positive selection at the molecular level because of its functional importance. Positive selection has been identified in the coding region in the human immunodeficiency virus (HIV)-1 envelope gene (Bonhoeffer *et al.* 1995), the major histocompatibility complex (MHC; Hughes and Nei 1988), the tumor suppressor gene BRCA1 (Huttley *et al.* 2000), female reproductive proteins in mammals (Swanson *et al.* 2001), and many other proteins (Yang and Bielawski 2000).

Several methods can be used to detect selection acting on protein-coding regions. One of the common approaches is to estimate the nonsynonymous rate (d_N) to synonymous rate (d_S) ratio (the most frequently used symbols in this article are given in Table 1). The d_N/d_S ratio is sometimes referred to as ω (*e.g.*, Goldman and Yang 1994). When a codon site undergoes negative selection, synonymous substitutions occur at a faster rate than nonsynonymous substitutions, and therefore $\omega < 1$. However, when there is no selection (neutrality), the rate of synonymous substitutions is equal to the rate of nonsynonymous substitutions, *i.e.*, $\omega = 1$. Alternatively, if the site undergoes positive selection, new mutations are beneficial and $\omega > 1$.

Inferences regarding ω have been used to demonstrate the presence of selection in many viral systems. Viruses may escape an existing immune response due to mutations in the proteins involved in interactions with the immune system. As a result several viral proteins have been observed to evolve under strong positive selection. In the HIV-1 envelope gene, positive selection has been found at sites that code for the surface positions in the protein (Bonhoeffer *et al.* 1995; Mindell 1996; Nielsen and Yang 1997; Yamaguchi and Gojobori 1998; Yamaguchi-Kabata and Gojobori 2000). In human influenza A (H3N2), positive selection has been found in the hemagglutinin (HA) gene, which encodes for a molecule that triggers the humoral immune response in humans (Fitch *et al.* 1997). Endo *et al.* (1996) and Yang and Bielawski (2000) gave more comprehensive lists of genes that are undergoing positive selection.

While positive selection has been found in the viral genes that code for proteins that interact with the host immune system, very little is known regarding selection in noncoding regions. A variety of research has shown that viral noncoding regions play an important role in gene regulation and function (Shiroki *et al.* 1995; Walker *et al.* 1995; Takayoshi *et al.* 1998). Furthermore, Carter and Roizman (1996) showed that viral introns are involved in alternative splicing and regulation of their own gene expression. The functional importance of the noncoding regions suggests that selection may be acting on them. However, since most interaction between the host immune system and viruses is at the level of proteins and peptides, very little positive selection is expected in noncoding regions compared to that in coding regions. Unfortunately, this and other hypotheses regarding selection in the noncoding region have not been subject to scientific evaluation because no appropriate statistical method has been available for this purpose.

We here extend the Nielsen and Yang (1998) and Yang *et al.* (2000) methods for detecting positive selection in coding regions to noncoding regions. We model the evolution in coding regions and the evolution in noncoding regions jointly and assume that the neutral

[1]*Corresponding author:* 434 Warren Hall, Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853. E-mail: sww8@cornell.edu

TABLE 1

Definitions of symbols used

Symbol	Definition
d_N	Number of nonsynonymous substitutions per nonsynonymous site
d_S	Number of synonymous substitutions per synonymous site
ω	Nonsynonymous/synonymous rate ratio (d_N/d_S)
κ	Transition/transversion rate ratio
$q_{ij}^{(C)}$	Rate of transition from codon i to codon j
$q_{ij}^{(N)}$	Rate of transition from nucleotide i to nucleotide j
π_j	Stationary frequency of codon j
μ_{j_k}	Stationary frequency of the nucleotide at position k in codon j
ζ	Nucleotide substitution rate multiplier in the noncoding region
n	Number of nucleotide sequences
N_N	Number of nucleotides in the noncoding region
N_C	Number of codons in the coding region
l	Log-likelihood

(synonymous) nucleotide substitution rate is constant in both the coding and noncoding regions of the same gene. Under this assumption, we introduce a new parameter ζ to model the evolution in the noncoding region. ζ is the nucleotide substitution rate in the noncoding region, normalized by the synonymous nucleotide substitution rate in the coding region. Therefore, when a site is subject to neutral selection, $\zeta = 1$. Similarly, $\zeta > 1$ indicates positive selection, while $\zeta < 1$ suggests the presence of negative selection. The interpretation of ζ is, therefore, similar to the interpretation of ω in models of evolution in coding regions. Using such a combined model we are able to develop a test to detect selection that is applicable to noncoding regions.

Three different simulated data sets are used to test the validity of the models. As an illustration of the method, we also compile 13 viral data sets from publications and GenBank to examine the hypothesis that positive selection mainly targets coding regions of viral genomes.

MODEL OF THE CODING REGION

We model the evolution of coding regions using a continuous-time Markov chain model of codon evolution. This model was proposed by Goldman and Yang (1994) and Muse and Gaut (1994). Like their models, the state space is given by the 61 sense codons, since the 3 stop codons are not allowed in the model. The rate matrix is 61×61, $Q^{(C)} = \{q_{ij}^{(C)}\}$, where $q_{ij}^{(C)}$ $(i \neq j)$ is the rate of transition from codon i to codon j. The transition rate from codon i to codon j is assumed to be proportional to the stationary distribution of the substituted nucleotide in codon j. If we represent codon i as a triplet $i_1 i_2 i_3$ and codon j as $j_1 j_2 j_3$ ($i_1, i_2, i_3, j_1, j_2, j_3 \in$ {T, C, A, G}), and if codons i and j differ by exactly one nucleotide in position k, then

$$q_{ij}^{(C)} = \begin{cases} \mu_{j_k} & \text{if } i_k \neq j_k \text{ by a synonymous transversion} \\ \kappa\mu_{j_k} & \text{if } i_k \neq j_k \text{ by a synonymous transition} \\ \omega\mu_{j_k} & \text{if } i_k \neq j_k \text{ by a nonsynonymous transversion} \\ \omega\kappa\mu_{j_k} & \text{if } i_k \neq j_k \text{ by a nonsynonymous transition.} \end{cases} \quad (1)$$

Additionally, $q_{ij}^{(C)} = 0$ if codons i and j differ in more than one position. The diagonal entries are defined as $q_{ii}^{(C)} = -\Sigma_{j \neq i} q_{ij}^{(C)}$ to fulfill the mathematical requirement that the row sums must equal 0. The parameter κ is the transition/transversion rate ratio. ω, as defined before, is the nonsynonymous/synonymous (d_N/d_S) rate ratio. μ_{j_k} is the stationary distribution of j_k, the nucleotide in the kth position of the codon. For instance, $q_{\text{CAC,CAG}}^{(C)} = \omega\pi_G$, since it is a nonsynonymous transversion (from histidine to glutamine and C $\rightarrow$ G is a transversion). The stationary frequency of codon i ($i_1 i_2 i_3$) is assumed to be the product of the stationary frequencies of nucleotides i_1, i_2, and i_3, divided by the sum of the stationary frequencies of the sense codons:

$$\pi_i = \frac{\mu_{i_1} \cdot \mu_{i_2} \cdot \mu_{i_3}}{c}, \quad \text{where } c = 1 - \pi_{TAA} - \pi_{TAG} - \pi_{TGA}. \quad (2)$$

Here we assume a universal genetic code, but with slight modification the method can be applied to other genetic codes.

To reduce the number of free parameters, we restrict our model to be time reversible, *i.e.*, $\pi_i q_{ij}^{(C)} = \pi_j q_{ji}^{(C)}$ for any i, j. It is easy to prove that this is indeed the case using standard methods. Furthermore, we take advantage of Felsenstein's (1981) pruning algorithm to save computational time.

Note that this model is different from the Goldman and Yang (1994) model. In our model the codon transition substitution rates depend on the stationary nucleotide frequencies, whereas in the Goldman and Yang (1994) model, the codon transition rates depend on

the stationary codon frequencies. κ and the stationary nucleotide substitution rates (μ_T, μ_C, μ_A, μ_G) govern the neutral mutation rates while ω models the effect of selection.

MODEL OF THE NONCODING REGION

The Hasegawa-Kishino-Yano (HKY)85 model (Hasegawa *et al.* 1985) is used in the noncoding region, with a parameterization that allows comparisons of substitution rates between coding and noncoding regions. Here, the unit of evolution is one nucleotide. Substitutions between nucleotides are described by a continuous-time Markov process. The instantaneous substitution rate from nucleotide i to j ($i \neq j$) is given by

$$q_{ij}^{(N)} = \begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{bmatrix} \cdot & \kappa\mu_C\zeta & \mu_A\zeta & \mu_G\zeta \\ \kappa\mu_T\zeta & \cdot & \mu_A\zeta & \mu_G\zeta \\ \mu_T\zeta & \mu_C\zeta & \cdot & \kappa\mu_G\zeta \\ \mu_T\zeta & \mu_C\zeta & \kappa\mu_A\zeta & \cdot \end{bmatrix} \end{array}, \tag{3}$$

where μ_i is the equilibrium frequency of nucleotide i, where $i \in \{A, C, T, G\}$. ξ is the rate of substitution relative to the neutral selection rate. As in the codon model, the diagonal entries are $q_{ii}^{(N)} = -\Sigma_{j \neq i} q_{ij}^{(N)}$.

Under this model, the estimate of ζ at a site indicates the type of selection acting on it. As in the models by Gaut and Weir (1994), Muse (1995), Nielsen and Yang (1998), and Yang *et al.* (2000) and in similar work by other authors, we assume that the synonymous substitution rate in the coding region reflects the neutral nucleotide substitution rate. Then

$$\zeta \begin{cases} < 1 & \text{if the site undergoes negative selection} \\ = 1 & \text{if the site undergoes neutral selection} \\ > 1 & \text{if the site undergoes positive selection.} \end{cases} \tag{4}$$

When $\zeta = 1$, the rate of nucleotide substitution at a site is equal to the synonymous codon substitution rate in the coding region, which in turn equals the synonymous nucleotide substitution rate. To illustrate this, consider C → G changes in both the noncoding and coding regions, and assume that the C → G change in the coding region is a change of codon TCC (Serine) to TCG (Serine). In the noncoding region, $q_{C,G}^{(N)} = \zeta\mu_G$ since the C → G change is a transversion, whereas in the coding region, $q_{TCC,TCG}^{(C)} = \mu_G$ since the change is a synonymous transversion. Likewise, $\zeta > 1$ implies that the substitution rate is greater than the neutral nucleotide substitution rate, whereas $\zeta < 1$ implies that the nucleotide substitution rate is lower than the neutral rate.

In our study we consider three rate classes of sites (negative, neutral, and positively evolving) in the noncoding region. We have implemented three models, namely the neutral model, the two-category model, and the three-category model (see Table 2). In the neutral model, we assume that there is no positive selection. Therefore, ζ can only be ≤ 1. In the two-category model, ζ can be <1, and it can be ≥ 1. In the three-category model, ζ can take on values in three categories: $\zeta < 1$, $\zeta = 1$, and $\zeta > 1$; that is,

$$\zeta = \begin{cases} \zeta_0 & \text{with probability } p_0 \\ \zeta_1 & \text{with probability } p_1 \\ \zeta_2 & \text{with probability } p_2, \end{cases}$$

where $0 < \zeta_0 < 1$, $\zeta_1 = 1$, $1 < \zeta_2 < \infty$, and $p_1 + p_2 + p_3 = 1$.

The neutral model is a special case of the three-category model in which $p_2 = 0$ and a special case of the two-category model in which $\zeta_2 = 1$. Since the neutral model is nested within the two other models, a likelihood-ratio test of the hypothesis of no positive selection can be performed by comparing the maximum-likelihood value obtained under the two- and three-category models to the maximum-likelihood value obtained under the neutral model.

COMBINING BOTH THE NONCODING AND CODING REGIONS

We assume that the DNA sequences are related through a single phylogenetic tree with a known topology, and we assume no recombination within each region analyzed. The true topology of the tree is rarely known; however, the codon-based likelihood methods for detecting positive selection have been shown to be highly robust to the assumptions regarding the topology of the phylogenetic tree (Yang *et al.* 2000). In practice, a good estimate of the topology of the tree will suffice. Joint optimization of the continuous parameters and the tree topology is currently not computationally feasible for the codon-based models.

In our joint model of coding and noncoding regions, the individual sites (codon sites in the case of the coding region and nucleotide sites in the case of the noncoding region) are assumed to be independent. Therefore, given the model and a particular phylogenetic tree, the log likelihood can be calculated as the sum of the log likelihoods among sites,

$$\begin{aligned} l = &\sum_{h^{(N)}=1}^{N_N} \log\{f(x_h|t_1, t_2, \ldots, \zeta_1, \zeta_2, \zeta_3, p_1, p_2, p_3, \kappa, \pi_T, \pi_C, \pi_A, \pi_G)\} \\ &+ \sum_{h^{(C)}=1}^{N_C} \log\{f(x_h|t_1, t_2, \ldots, \omega, \kappa, \pi_T, \pi_C, \pi_A, \pi_G)\}, \end{aligned} \tag{5}$$

where N_N is the number of nucleotides in the noncoding region, N_C is the number of codons in the coding region, and t_1, t_2 . . . are the branch lengths of the tree.

The log likelihood defined in Equation 5 can be optimized using standard optimization techniques. The popular local unconstrained optimization procedure [Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm] adopted from *Numerical Recipes in C* (Press *et al.* 1992)

TABLE 2

Models of variable ζ among sites

Model	P^a	Estimable parameters	Constraints
Neutral	2	ζ_0, p_0	$\zeta_0 < 1$, $\zeta_1 = 1$, $p_1 = 1 - p_0$
Two-category	3	ζ_0, ζ_1, p_0	$\zeta_0 < 1$, $\zeta_1 \geq 1$, $p_1 = 1 - p_0$
Three-category	4	ζ_0, ζ_2, p_0, p_1	$\zeta_0 < 1$, $\zeta_1 = 1$, $\zeta_2 > 1$, $p_2 = 1 - p_0 - p_1$

[a] P, number of estimable parameters in the ζ distribution.

is used in our program EvoNC. However, since the parameters in our models are constrained (all parameters are constrained to be >0, the proportions of ζ in each category have to add 1 to 1, and some ζ's have to be <1 and some >1), it is a constrained optimization problem. We replaced the original objective function by a quadratic penalty function. The quadratic penalty function consists of the original objective function and a penalty term for each constraint. The penalty term is a multiple of the square of the constraint violation when the current parameter vector violates the constraint and is 0 otherwise (Nocedal and Wright 1999). A barrier term was also incorporated for each parameter to ensure that the vector of parameters lies in the interior of the parameter space.

In Nielsen and Yang (1998) and Yang *et al.* (2000) ω was allowed to vary among sites. In this study we do not pursue such models as our primary interest is in the noncoding region. Allowing ω to vary among sites in the codon region is likely to increase the likelihood of the model. It might affect the parameter estimates of ζ since ζ and ω may be correlated. However, implementing a model that allows variable ω will make the optimization procedure even more computationally intensive.

LIKELIHOOD-RATIO TESTS AND POSTERIOR DECODING

In general, if we have a model with k categories of ζ, let the corresponding proportions of nucleotide sites in the three categories be $p_1 \ldots p_k$, under the constraint that $p_1 + \ldots + p_k = 1$. The probability of observing data (x_h) at site h is then

$$f(x_h) = \sum_{i=1}^{k} p_i f(x_h|\zeta_i). \quad (6)$$

The conditional probability can be calculated using an empirical Bayes approach as shown in Nielsen and Yang (1998). The posterior probability that the nucleotide site h belongs to category k is given by

$$\text{prob}(\zeta_i|x_h) = \frac{p_i f(x_h|\zeta_i)}{f(x_h)} = \frac{p_i f(x_h|\zeta_i)}{\sum_{l=1}^{k} p_l f(x_h|\zeta_l)}. \quad (7)$$

We assign sites to categories by choosing the category k that maximizes $\text{prob}(\zeta_k|x_h)$.

The maximum-likelihood framework allows us to test the null hypothesis of no positive selection using likelihood-ratio tests. Two different likelihood-ratio tests can be performed by comparing the neutral model against the three-category model and the neutral model against the two-category model. The three-category model has two more parameters (ζ_3, the category that allows $\zeta > 1$, and p_3, the proportion of ζ_3) than the neutral model. Comparing twice the log-likelihood difference between these two models with the χ^2 distribution with 2 d.f. may be used to approximate P values of this test. However, because one of the parameters is on the boundary of the parameter space and another parameter is not estimable under the null hypothesis, the true asymptotic distribution of the likelihood-ratio test statistic is not known under the null hypothesis. The resulting test will therefore be conservative. A better test, similar to test II in Swanson *et al.* (2003) for the coding region, is to compare the neutral model against the two-category model. In this test twice the difference in log-likelihood is asymptotically distributed as a 50:50 mixture of a point mass at 0 and a χ^2 distribution with 1 d.f. (Chernoff 1954; Self and Liang 1987). The benefits of this test are that the reduction in the degrees of freedom may in some cases increase the power of the test and that the true asymptotic distribution of the test statistic is actually known.

SIMULATION

Nucleotide sequence data were simulated using Evolver in the PAML package (Yang 1997) to verify EvoNC. Each simulated data set consisted of 15 sequences of 1400 nucleotides. Each sequence was composed of 200 noncoding nucleotide sites and 400 codon sites (1200 nucleotides). In the noncoding region, positive, neutral, and negative selected data were simulated by varying the branch lengths. In Evolver, the tree length specifies the expected number of substitutions per site along the branches in the tree. Therefore, if we let the neutral sites have a total tree length equal to 1, then the sites that evolve five times faster would have the total tree length equal to 5. Similarly, the sites that evolve five times slower can be simulated by letting the total tree length be 0.2. In the coding region, ω was set to be 0.4. Both regions shared the same κ = 5. The

TABLE 3

Percentage of ζ of the three categories in the four simulation data sets

	$\zeta = 0.1$	$\zeta = 1.0$	$\zeta = 2$	$\zeta = 5$
Neutral set	0	100	0	0
Conserved set	50	50	0	0
Positive set 1	72.5	25	0	2.5
Positive set 2	72.5	25	2.5	0

The total number of sites in the noncoding region is 200 in each set.

proportions of the three categories of ζ were chosen with an intention to reflect the distribution in real data sets. The distribution of ζ in the four data sets is shown in Table 3.

DATA ANALYSIS

The viral data sets were assembled by searching Gen Bank for previously published data sets. We used Clustal W version 1.8 (Thompson *et al.* 1994) for multiple alignments of the data sets. To prevent false positive results because of misalignments, data with poor alignments and/or with detectable recombination were discarded. For the same reason, because interspecific viral sequences are often too diverged to give a reasonably good alignment, only intraspecific viral data were used. We also made sure that the data did not consist of published results of overlapping reading frames.

Table 4 gives the summarized source and the details of the 13 data sets and the GenBank accession numbers are available from the authors upon request. NEIGHBOR in the PHYLIP package version 3.6 (Felsenstein 2002) was used for estimating phylogenetic trees. The alignment gaps were removed by our program EvoNC.

The coding region of each of the data sets was analyzed by codeml in the PAML package version 3.12 (Yang 1997). Models M7, M8 (Yang *et al.* 2000), and M8a (Swanson *et al.* 2003) were used in the study to determine if positive selection exists in the coding region.

For each of the data sets, two tests were performed for the coding region. Test 1 compares twice the log-likelihood ratio difference between M7 and M8 to a χ^2 distribution with 2 d.f.. As noted in Yang *et al.* (2000), this test is conservative. Test 2 compares twice the log-likelihood difference between M8 and M8a to a $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ distribution, as suggested in Swanson *et al.* (2003).

The selection in the noncoding region was investigated using EvoNC. Both the coding and the noncoding regions were used in the program. Similar to the coding region, two tests were performed for each of the data sets. Test 1 compares twice the log-likelihood ratios between the neutral and the three-category model to a χ_2^2 distribution, whereas test 2 compares twice the log-likelihood ratios between the neutral and two-category model to a $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ distribution. The critical value of the likelihood-ratio test statistic, for a test performed at the 5% significance level, is then 2.71.

RESULTS

Simulation data: The simulated data were analyzed using the three models (the neutral model, the two-category model, and the three-category model, as mentioned above), and the log-likelihood of each model was used to perform likelihood-ratio tests. Estimates of all parameters of the model were obtained from the three-category model. Each site was categorized according to the posterior probabilities calculated using Equation 8. The estimates of parameters are summa-

TABLE 4

Details of the 13 viral data sets

Virus[a]	n	N_N[b]	N_C	Reference
Hog cholera virus polyprotein	8	182	42	Vilcek and Belak (1997)
Dengue virus type 1	9	463	3392	Direct GenBank search
Ebola glycoprotein	5	(3′) 245	688	Sanchez *et al.* (1996)
Foot-and-mouth disease	5	110	2332	Direct GenBank search
Hepatitis A	6	734	2227	Fujiwara *et al.* (2001)
Hepatitis C	9	(3′) 344	3043	Salemi and Vandamme (2002)
Japanese encephalitis virus	20	(3′) 586	195	Direct GenBank search
Mammalian orthoreovirus	7	12	1289	Breun *et al.* (2001)
Newcastle disease virus nucleocapsid protein	9	65	489	Seal *et al.* (2002)
Poliovirus	10	743	2209	Direct GenBank search
Rabies virus glycoprotein	35	497	524	Badrane and Tordo (2001)
TT virus ORF 2	8	171	232	Luo *et al.* (2002)
West Nile virus	12	95	3434	Direct GenBank search

[a] Complete genome of the virus unless specified otherwise.
[b] 5′ noncoding region unless specified otherwise.

TABLE 5

True and estimated values of parameters of the simulated data under the three-category model

	True values of the parameters			Estimates of parameters		
Data set	κ	d_N/d_S	ζ's and *p*'s	κ	d_N/d_S	ζ's and *p*'s
Neutral	5.00	0.40	ζ_0 = NE[a], $p_0 = 0.00$, $p_1 = 1.00$, ζ_2 = NE, ($p_2 = 0.00$)	4.59	0.32	$\zeta_0 = 1.00$, $p_0 = 1.00$, $p_1 = 0.00$, $\zeta_2 = 1.00$, ($p_2 = 0.00$)
Conserved	5.00	0.40	$\zeta_0 = 0.20$, $p_0 = 0.50$, $p_1 = 0.50$, ζ_2 = NE, ($p_2 = 0.00$)	5.02	0.40	$\zeta_0 = 0.20$, $p_0 = 0.51$, $p_1 = 0.26$, $\zeta_2 = 1.00$, ($p_2 = 0.23$)
Positive 1	5.00	0.40	$\zeta_0 = 0.20$, $p_0 = 0.72.5$, $p_1 = 0.25$, $\zeta_2 = 5$, $p_2 = 0.025$	4.73	0.39	$\zeta_0 = 0.23$, $p_0 = 0.79$, $p_1 = 0.15$, $\zeta_2 = 4.39$, ($p_2 = 0.06$)
Positive 2	5.00	0.40	$\zeta_0 = 0.20$, $p_0 = 0.72.5$, $p_1 = 0.25$, $\zeta_2 = 2$, $p_2 = 0.025$	3.86	0.36	$\zeta_0 = 0.21$, $p_0 = 0.79$, $p_1 = 0.11$, $\zeta_2 = 1.55$, ($p_2 = 0.10$)

[a] NE, not estimable.

rized in Table 5, and the classification of sites in conserved and positive sets is shown in Figure 1.

The estimated values of ζ match quite well with the actual values. In the neutral data set, estimates of $\hat{\zeta}_0 = \hat{\zeta}_2 = \hat{\zeta}_1 = 1$ were obtained for all sites for all models. In this boundary of the parameter space, the proportion of sites in the different categories is not identifiable. In the conserved data set, maximum-likelihood estimates of $\hat{\zeta}_0 = 0.20$ and $\hat{\zeta}_2 = \hat{\zeta}_1 = 1.00$ were obtained for the three-category model, which were virtually identical to the true values of 0.2 and 1.0. Again, the proportion of sites in category 1 and 2 is not identifiable. The estimate of the proportion of conserved sites was $\hat{p}_0 = 0.51$, which was slightly larger than the true proportion 0.50. In the positive data set 1, estimates of $\hat{\zeta}_0 = 0.23$ and $\hat{\zeta}_2 = 4.38$ were obtained for the three-category model, which differed slightly from the true values of 0.2 and 5.0, respectively. The estimated proportion of neutral sites was less than expected (0.15 *vs.* 0.25). On the other hand, $\hat{p}_0 = 0.79$ and $\hat{p}_2 = 0.06$, which were both larger than the true values of 0.725 and 0.025. In the positive data set 2, the estimates from the three-category model are $\hat{\zeta}_0 = 0.21$ and $\hat{\zeta}_2 = 1.55$, and the corresponding proportions are $\hat{p}_0 = 0.79$ and $\hat{p}_2 = 0.10$. Again, the estimated proportions were slightly larger than the true values.

In the neutral and conserved data sets, all three models gave approximately the same maximum-log-likelihood value in these two sets. In these two cases, the neutral null hypothesis was true and was not rejected by the likelihood-ratio test.

In the positive data set 1, the maximum-log-likelihood values for the three models were

$$l = \begin{cases} -13252.97 & \text{in the neutral model} \\ -13249.15 & \text{in the two-category model} \\ -13245.77 & \text{in the three-category model.} \end{cases}$$

Both tests reject the false null hypothesis of no positive selection. However, from the maximum-log-likelihood shown below, only test 2 (neutral model *vs.* two-category model) was significant in positive data set 2. This is probably due to the fact that only 5 of 200 sites were slightly positively selected.

$$l = \begin{cases} -10285.83 & \text{in the neutral model} \\ -10284.26 & \text{in the two-category model} \\ -10284.02 & \text{in the three-category model.} \end{cases}$$

The Bayesian classification of sites performed quite well for the first three simulated data sets. In the neutral data set, EvoNC classified all 200 sites correctly. In the conserved set, 14 of the actual neutral sites were miscategorized as conserved sites while 3 of the actual conserved sites were miscategorized as neutral sites. A total of 183 sites were classified correctly. In positive data set 1, the 5 positively selected sites were all included in the estimated set of positively selected sites. Three neutral sites were also falsely included in this set.

In positive data set 2, 1 out of the 5 positively selected sites was classified as being neutral and 16 out of the 195 negatively selected and neutral sites were classified as being positively selected. The main reason the method performs worse when there are only a few positively selected sites is that the estimates of $\hat{\zeta}_2$ are more unreliable. When the maximum-likelihood estimates of parameters have large variance, the empirical Bayes estimates of posterior probabilities may have reduced accuracy. Nevertheless, a close examination revealed that the falsely classified sites have low posterior probabilities. Therefore, by adjusting the cutoff probability the accuracy of the method can be controlled.

These results suggest that our method is capable of picking up strong positive selection, even though as few as 2.5% of the sites are positively selected. On the other hand, when only a small portion of the sites are undergoing weak positive selection, the classification of sites does not have great accuracy.

The viral data sets: The results of the PAML analysis are summarized in Table 6. Five out of the 13 viral data sets have significant evidence for positive selection, using test 1 for the coding region (M7 *vs.* M8): glycoprotein of Ebola virus, foot-and-mouth disease polyprotein, hepatitis C polyprotein, New Castle disease virus nucleo-

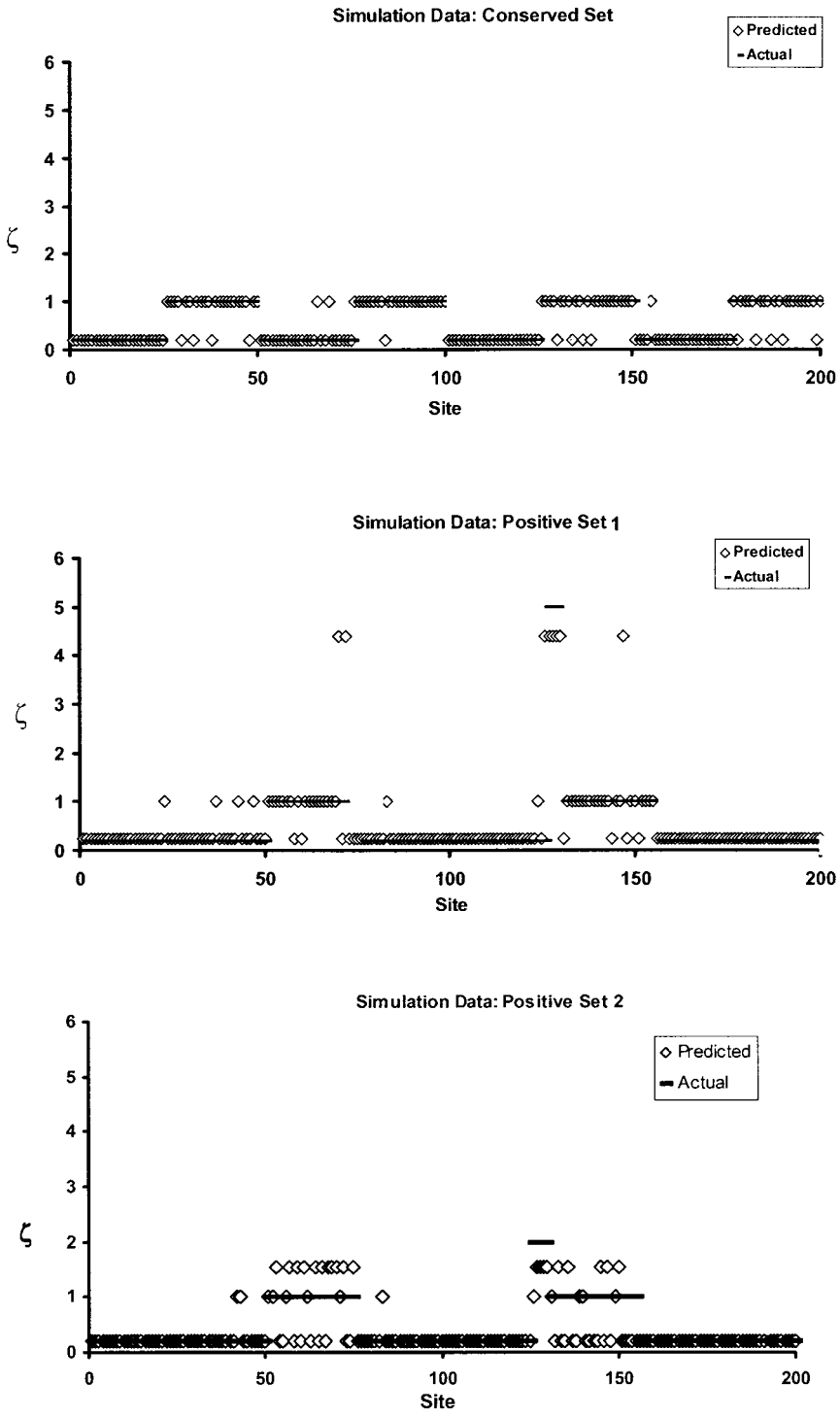

FIGURE 1.—Predicted *vs.* actual values of ζ in two of the simulated data sets. The three-category model was used for prediction. Predicted parameter values are the MAP (maximum *a priori*) estimates.

capsid protein, and poliovirus polyprotein. Among these 5 data sets, 2 were not significant using test 2 for the coding region (M8a *vs.* M8). This may be because the true value of ω was only slightly >1 in the positively selected sites or, possibly, because the beta distribution assumed in model M7 did not fit the true distribution of ω well. For instance, in the foot-and-mouth disease data set, ω was equal to 1.01 in the positively selected sites. Clearly, this cannot be interpreted as good evidence for positive selection in this virus. On the other hand, for example, 11% of the sites in Ebola glycoprotein have ω = 5.92, which is probably a good indication that this protein is undergoing positive selection.

In the noncoding regions there was little or no evidence of positive selection (Table 6). For most data sets, the log-likelihoods of all three models were almost exactly the same (Table 7). One exception was the Japanese encephalitis virus. The test of the two-category

TABLE 6

Log-likelihood values under models M7, M8 (YANG *et al.* 2000), and M8a (SWANSON *et al.* 2003) and parameter estimates under model M8 by PAML in the coding region

Data set	Log-likelihood values of the models			Estimates of parameters p, q, and ω in M8
	M7	M8	M8a	
Classical swine fever virus	−365.60	−365.47	−365.60	$p_0 = 0.53$, $p = 0.00$, $q = 1.52$, $(p_1 = 0.47)$, $\omega = 0.59$
Dengue virus type 1	−23921.15	−23919.40	−23919.87	$p_0 = 0.96$, $p = 3.06$, $q = 99.00$, $(p_1 = 0.04)$, $\omega = 0.69$
Ebola glycoprotein	−8192.94	−8188.12	−8189.07	$p_0 = 0.89$, $p = 0.49$, $q = 7.58$, $(p_1 = 0.11)$, $\omega = 5.92$
Foot-and-mouth disease	−19071.89	−19066.11	−19066.11	$p_0 = 0.97$, $p = 0.13$, $q = 2.34$, $(p_1 = 0.03)$, $\omega = 1.01$
Hepatitis A	−11747.53	−11746.19	−11746.20	$p_0 = 0.99$, $p = 1.78$, $q = 99.00$, $(p_1 = 0.01)$, $\omega = 1.19$
Hepatitis C	−59726.45	−59717.34	−59720.92	$p_0 = 0.99$, $p = 0.27$, $q = 2.36$, $(p_1 = 0.01)$, $\omega = 73.73$
Japanese encephalitis virus	−28710.23	−28708.87	−28708.88	$p_0 = 0.56$, $p = 0.09$, $q = 0.92$, $(p_1 = 0.44)$, $\omega = 0.06$
Mammalian orthoreovirus	−13060.29	−13058.67	−13058.85	$p_0 = 0.64$, $p = 0.02$, $q = 0.42$, $(p_1 = 0.36)$, $\omega = 0.01$
Newcastle disease virus	−4787.83	−4776.37	−4785.65	$p_0 = 1.00$, $p = 0.23$, $q = 1.59$, $(p_1 = 0.00)$, $\omega = 99.00$
Poliovirus	−19803.71	−19799.22	−19799.50	$p_0 = 0.96$, $p = 0.30$, $q = 2.83$, $(p_1 = 0.04)$
Rabies virus glycoprotein	−11499.48	−11497.40	−11497.70	$p_0 = 0.98$, $p = 0.33$, $q = 3.88$, $(p_1 = 0.02)$, $\omega = 0.69$
TT virus ORF 2	−3666.58	−3666.29	−3666.35	$p_0 = 0.84$, $p = 0.80$, $q = 0.99$, $(p_1 = 0.16)$, $\omega = 1.13$
West Nile virus	−22769.51	−22769.33	−22769.51	$p_0 = 0.95$, $p = 0.62$, $q = 99.00$, $(p_1 = 0.05)$, $\omega = 0.14$

model *vs.* the neutral model in the 3′-untranslated region of Japanese encephalitis virus was marginally significant ($P = 0.05$), but the test of the three-category model *vs.* the neutral model ($P = 0.22$) was not. NAM *et al.* (2002) described this region as the variable region since it showed a high degree of sequence variation and deletion. However, they also pointed out that despite the fact that the region was highly variable, the predicted RNA structures all had a similar type loop at the 5′ terminus.

After correcting for multiple testing by the Bonferroni procedure, we conclude that the likelihood-ratio tests provide no evidence for positive selection in the viral data that we examined. This result confirms our expectation that positive selection occurs primarily in the coding regions of viruses.

DISCUSSION

Because positive selection is expected to occur in regions where viruses interact with the host immune system, this type of selection is expected to be much more predominant in coding than in noncoding regions. Our study confirmed this belief. However, it should be noted that most of the data sets in the study had a low number of sequences and perhaps had low sequence diversity as well. ANISIMOVA *et al.* (2001) showed that such data sets would reduce the power of the tests. Therefore, data sets with a low number of sequences should be reexamined when more data become available.

The results of our analysis of simulated data suggest that the new method provides accurate parameter estimates. The likelihood-ratio tests also performed well and detected selection from 15 sequences when only 2.5% of the sites were undergoing positive selection. When more than one category of ζ was present, the program miscategorized ~10% of the sites. In small data sets the classification of sites may not have the highest accuracy. A similar conclusion has been reached regarding classification of sites in coding regions (ANISIMOVA *et al.* 2001, 2002). In the analysis of real data, it is advisable to confirm the presence of positive selection in particular sites by employing additional structural, functional, or evolutionary information.

Our model is based on the assumption that there is no selection acting on the synonymous sites and the rate of substitution is constant among sites. This assumption

TABLE 7

Log-likelihood values and parameter estimates of the 13 viral data sets under the three models

				Estimates from the three-category model	
Data set	l_1	l_2	l_3	ζ's and p's	ζ
Classical swine fever virus	−832.27	−832.27	−832.27	$\zeta_0 = 0.02$, $p_0 = 0.59$, $p_1 = 0.41$, $\zeta_2 = 1.00$, ($p_2 = 0.00$)	0.43
Dengue virus type 1	−25194.86	−25193.90	−25193.81	$\zeta_0 = 0.00$, $p_0 = 0.77$, $p_1 = 0.12$, $\zeta_2 = 1.00$, ($p_2 = 0.11$)	0.23
Ebola glycoprotein	−9281.60	−9281.60	−9281.60	$\zeta_0 = 0.21$, $p_0 = 0.58$, $p_1 = 0.31$, $\zeta_2 = 1.00$, ($p_2 = 0.11$)	0.55
Foot-and-mouth disease	−19807.91	−19807.91	−19807.91	$\zeta_0 = 0.12$, $p_0 = 0.78$, $p_1 = 0.15$, $\zeta_2 = 1.00$, ($p_2 = 0.07$)	0.32
Hepatitis A	−13293.50	−13293.10	−13293.10	$\zeta_0 = 0.00$, $p_0 = 0.70$, $p_1 = 0.15$, $\zeta_2 = 1.81$, ($p_2 = 0.15$)	0.42
Hepatitis C	−62243.81	−62243.81	−62243.81	$\zeta_0 = 0.01$, $p_0 = 0.98$, $p_1 = 0.01$, $\zeta_2 = 1.00$, ($p_2 = 0.01$)	0.03
Japanese encephalitis virus	−30207.91	−30206.55	−30206.40	$\zeta_0 = 0.08$, $p_0 = 0.92$, $p_1 = 0.03$, $\zeta_2 = 2.38$, ($p_2 = 0.04$)	0.21
Mammalian orthoreovius	−13413.08	−13413.08	−13413.08	$\zeta_0 = 0.05$, $p_0 = 1.00$, $p_1 = 0.00$, $\zeta_2 = 57.45$, ($p_2 = 0.00$)	0.05
Newcastle disease virus	−5006.73	−5006.71	−5006.71	$\zeta_0 = 0.07$, $p_0 = 0.87$, $p_1 = 0.06$, $\zeta_2 = 1.00$, ($p_2 = 0.07$)	0.19
Poliovirus	−21857.32	−21857.32	−21857.32	$\zeta_0 = 0.00$, $p_0 = 0.91$, $p_1 = 0.05$, $\zeta_2 = 1.00$, ($p_2 = 0.04$)	0.14
Rabies virus glycoprotein	−17398.91	−17398.49	−17398.49	$\zeta_0 = 0.30$, $p_0 = 0.53$, $p_1 = 0.34$, $\zeta_2 = 1.00$, ($p_2 = 0.13$)	0.63
TT virus ORF 2	−3517.02	−3517.02	−3517.02	$\zeta_0 = 0.01$, $p_0 = 0.91$, $p_1 = 0.05$, $\zeta_2 = 1.00$, ($p_2 = 0.04$)	0.10
West Nile virus	−23067.52	−23067.49	−23067.38	$\zeta_0 = 0.00$, $p_0 = 0.98$, $p_1 = 0.01$, $\zeta_2 = 2.95$, ($p_2 = 0.01$)	0.65

would be violated if there is codon usage bias in the coding region. A lot of the viruses may have codon usage bias due to the overlapping reading frames and/or RNA structural constraints in coding regions. We do not know how much the bias would have affected the parameter estimates but this is a question that could be addressed using simulations.

The models presented here do not allow rate variation among the lineages in the phylogeny. This may result in a loss of statistical power of the likelihood-ratio tests. Our method will not be able to identify positively selected sites if positive selection occurs in only a few lineages, while the negative selection dominates the rest. In the future, we would like to incorporate rate variation among the lineages into the program as well.

Since the optimization procedure (BFGS algorithm) used here is a local optimization procedure, it is possible that the likelihood is trapped at a local minimum. While we do not have a rigorous solution to the problem, it is advisable to run EvoNC with different sets of initial values. If they all result in the same log-likelihood it is very likely that one has found the global minimum.

EvoNC is currently under development for incorporating new models and functionalities. The beta version for the models tested in this article is available from the authors upon request.

We are grateful to Charles F. Aquadro and David B. Shmoys for their valuable comments. This work was supported by National Science Foundation (NSF) grant DEB-0089487, NSF/National Institutes of Health grant DMS/NIGMS-0201037, and Human Frontier in Science Program grant RGY0055/2001-M.

LITERATURE CITED

Anisimova, M., J. P. Bielawski and Z. Yang, 2001 Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. **18:** 1585–1592.

Anisimova, M., J. P. Bielawski and Z. Yang, 2002 Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol. Biol. Evol. **19:** 950–958.

Badrane, H., and N. Tordo, 2001 Host switching in Lyssavirus history from the Chiroptera to the Carnivora orders. J. Virol. **75:** 8096–8104.

Bonhoeffer, S., E. C. Holmes and M. A. Nowak, 1995 Causes of HIV diversity. Nature **376:** 125.

Breun, L. A., T. J. Broering, A. M. McCutcheon, S. J. Harrison, C. L. Luongo *et al.*, 2001 Mammalian reovirus L2 gene and lambda2 core spike protein sequences and whole-genome comparisons of reoviruses type 1 Lang, type 2 Jones, and type 3 Dearing. Virology **287:** 333–348.

Carter, K. L., and B. Roizman, 1996 Alternatively spliced mRNAs predicted to yield frame-shift proteins and stable intron 1 RNAs of the herpes simplex virus 1 regulatory gene alpha 0 accumulate in the cytoplasm of infected cells. Proc. Natl. Acad. Sci. USA **93:** 12535–12540.

Chernoff, H., 1954 On the distribution of the likelihood ratio. Ann. Math. Stat. **25:** 573–578.

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.

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

Felsenstein, J., 2002 *Phylogenetic Inference Package (PHYLIP), Version 3.6.* University of Washington, Seattle.

Fitch, W. M., R. M. Bush, C. A. Bender and N. J. Cox, 1997 Long term trends in the evolution of H(3) HA1 human influenza type A. Proc. Natl. Acad. Sci. USA **94:** 7712–7718.

Fujiwara, K., O. Yokosuka, K. Fukai, F. Imazeki, H. Saisho *et al.*, 2001 Analysis of full-length hepatitis A virus genome in sera from patients with fulminant and self-limited acute type A hepatitis. J. Hepatol. **35:** 112–119.

Gaut, B. S., and B. S. Weir, 1994 Detecting substitution-rate heterogeneity among regions of a nucleotide sequence. Mol. Biol. Evol. **11:** 620–629.

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

Hasegawa, M., H. Kishino and T. Yano, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22:** 160–174.

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.

Huttley, G. A., S. Easteal, M. C. Southey, A. Tesoriero, G. G. Giles *et al.*, 2000 Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Australian Breast Cancer Family Study. Nat. Genet. **25:** 410–413.

Luo, K., H. He, Z. Liu, D. Liu, H. Xiao *et al.*, 2002 Novel variants related to TT virus distributed widely in China. J. Med. Virol. **67:** 118–126.

Mindell, D. P., 1996 Positive selection and rates of evolution in immunodeficiency viruses from humans and chimpanzees. Proc. Natl. Acad. Sci. USA **93:** 3284–3288.

Muse, S. V., 1995 Evolutionary analyses of DNA sequences subject to constraints of secondary structure. Genetics **139:** 1429–1439.

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.

Nam, J. H., S. L. Chae, S. H. Park, Y. S. Jeong, M. S. Joo *et al.*, 2002 High level of sequence variation in the 3′ noncoding region of Japanese encephalitis viruses isolated in Korea. Virus Genes **24:** 21–27.

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.

Nocedal, J., and S. J. Wright, 1999 *Numerical Optimization*, pp. 490–527. Springer-Verlag, New York.

Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, 1992 *Numerical Recipes: The Art of Scientific Computing*, pp. 425–430. Cambridge University Press, Cambridge/London/New York.

Salemi, M., and A. M. Vandamme, 2002 Hepatitis C virus evolutionary patterns studied through analysis of full-genome sequences. J. Mol. Evol. **54:** 62–70.

Sanchez, A., S. G. Trappier, B. W. Mahy, C. J. Peters and S. T. Nichol, 1996 The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing. Proc. Natl. Acad. Sci. USA **93:** 3602–3607.

Seal, B. S., J. M. Crawford, H. S. Sellers, D. P. Locke and D. J. King, 2002 Nucleotide sequence analysis of the Newcastle disease virus nucleocapsid protein gene and phylogenetic relationships among the Paramyxoviridae. Virus Res. **83:** 119–129.

Self, S., and K.-Y. Liang, 1987 Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. **82:** 605–610.

Shiroki, K., T. Ishii, T. Aoki, M. Kobashi, S. Ohka *et al.*, 1995 A new cis-acting element for RNA replication within the 5′ noncoding region of poliovirus type 1 RNA. J. Virol. **69:** 6825–6832.

Swanson, W. J., Z. Yang, M. F. Wolfner and C. F. Aquadro, 2001 Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc. Natl. Acad. Sci. USA **98:** 2509–2514.

Swanson, W. J., R. Nielsen and Q. Yang, 2003 Pervasive adaptive evolution in mammalian fertilization proteins. Mol. Biol. Evol. **20:** 18–20.

Takayoshi, I., S. M. Tahara and M. M. C. Lai, 1998 The 39-untranslated region of hepatitis C virus RNA enhances translation from an internal ribosomal entry site. J. Virol. **72:** 8789–8796.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Vilcek, S., and S. Belak, 1997 Organization and diversity of the 3′-noncoding region of classical swine fever virus genome. Virus Genes **15:** 181–186.

Walker, P. A., L. E. Leong and A. G. Porter, 1995 Sequence and structural determinants of the interaction between the 5′-noncoding region of picornavirus RNA and rhinovirus protease 3C. J. Biol. Chem. **270:** 14510–14516.

Yamaguchi, Y., and T. Gojobori, 1997 Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. Proc. Natl. Acad. Sci. USA **94:** 1264–1269.

Yamaguchi-Kabata, Y., and T. Gojobori, 2000 Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. J. Virol. **74:** 4335–4350.

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

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. Pedersen, 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:** 431–449.

Communicating editor: J. J. Hein