Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites

Maria Anisimova,*,†,[1] Rasmus Nielsen‡ and Ziheng Yang*

**Department of Biology, University College London, London WC1E 6BT, United Kingdom, †Center for Mathematics and Physics in the Life Sciences and Experimental Biology (CoMPLEX), University College London, London WC1E 6BT, United Kingdom and ‡Department of Biometrics, Cornell University, Ithaca, New York 14853-7801*

Manuscript received July 30, 2002
Accepted for publication March 31, 2003

ABSTRACT

Maximum-likelihood methods based on models of codon substitution accounting for heterogeneous selective pressures across sites have proved to be powerful in detecting positive selection in protein-coding DNA sequences. Those methods are phylogeny based and do not account for the effects of recombination. When recombination occurs, such as in population data, no unique tree topology can describe the evolutionary history of the whole sequence. This violation of assumptions raises serious concerns about the likelihood method for detecting positive selection. Here we use computer simulation to evaluate the reliability of the likelihood-ratio test (LRT) for positive selection in the presence of recombination. We examine three tests based on different models of variable selective pressures among sites. Sequences are simulated using a coalescent model with recombination and analyzed using codon-based likelihood models ignoring recombination. We find that the LRT is robust to low levels of recombination (with fewer than three recombination events in the history of a sample of 10 sequences). However, at higher levels of recombination, the type I error rate can be as high as 90%, especially when the null model in the LRT is unrealistic, and the test often mistakes recombination as evidence for positive selection. The test that compares the more realistic models M7 (β) against M8 (β and ω) is more robust to recombination, where the null model M7 allows the positive selection pressure to vary between 0 and 1 (and so does not account for positive selection), and the alternative model M8 allows an additional discrete class with $\omega = d_N/d_S$ that could be estimated to be >1 (and thus accounts for positive selection). Identification of sites under positive selection by the empirical Bayes method appears to be less affected than the LRT by recombination.

ADAPTIVE molecular evolution has long been a subject of intense interest among evolutionary biologists. For protein-coding genes, robust evidence for positive selection is provided by an excess of nonsynonymous substitutions relative to synonymous substitutions (see Yang and Bielawski 2000 for review). If a change of amino acid offers a selective advantage, causing accelerated fixation of the nonsynonymous mutation, the nonsynonymous substitution rate d_N will be higher than the synonymous rate d_S, with the rate ratio $\omega = d_N/d_S > 1$. Since positive selection is not expected to act at all amino acid sites, much effort has been taken to account for variable selective pressures across sites to improve the power of the methods for detecting positive selection (*e.g.*, Fitch *et al.* 1997; Nielsen and Yang 1998; Suzuki and Gojobori 1999; Yamaguchi-Kabata and Gojobori 2000; Yang *et al.* 2000a). Such methods have little power to detect positive episodic or directional selection but have been successful in detecting recurrent diversifying selection. Likelihood-ratio tests (LRTs) proposed by Nielsen and Yang (1998) and Yang *et al.* (2000a) detect positive selection by comparing two nested probabilistic models of variable ω ratios among sites, the simpler of which does not allow for sites with $\omega > 1$ and the more general of which does. When the LRT suggests presence of sites under positive selection, the empirical Bayes approach can be used to identify locations of those sites in a sequence (Nielsen and Yang 1998). Although the underlying evolutionary process is almost certainly more complex than existing models, the maximum-likelihood (ML) approach provides a statistically sound framework for testing for presence of sites under positive selection, measuring the strength of selection, and identifying critical amino acids under selection (Yang *et al.* 2000a; Anisimova *et al.* 2002).

A number of genes have been detected by the LRT to be undergoing positive selection. Among the nonviral examples are mammalian β-globin, mitochondrial genes from hominoids (Yang *et al.* 2000a), plant chitinases (Bishop *et al.* 2000), abalone sperm lysin (Yang *et al.* 2000b), mammalian female reproductive proteins (Swanson *et al.* 2001), salmonid iron-binding proteins (Ford 2001), and fimbrial adhesins of *Escherichia coli* (Peek *et al.* 2001). Positive selective pressure was detected with LRTs in a number of viral genes: capsid genes of foot-and-mouth virus (Fares *et al.* 2001; Haydon *et al.* 2001),

[1]*Corresponding author:* Department of Biology, University College London, Darwin Bldg., Gower St., London WC1E 6BT, United Kingdom. E-mail: m.anisimova@ucl.ac.uk

TABLE 1

Models of ω-ratio variation among sites used for simulation and analysis

Model code	Description	Free parameters
M0: one ratio	Constant ω for all sites	ω
M1: neutral	p_0 of sites with $\omega_0 = 0$ p_1 of sites with $\omega_1 = 1$	p_0
M2: selection	Three site classes with $\omega_0 = 0$, $\omega_1 = 1$, ω_2 in proportions p_0, p_1, p_2	p_0, p_1, ω_2
M3: discrete	Three site classes in proportions p_0, p_1, p_2	p_0, p_1, ω_0, ω_1, ω_2
M7: β	All sites are from $B(p, q)$	p, q
M8: β and ω	p_0 of sites are from $B(p, q)$ p_1 of sites are from a class with ω	p, q, p_0, ω

the G and N genes of rabies virus (Holmes *et al.* 2002), and major HIV-1 genes (Nielsen and Yang 1998; Yang *et al.* 2000a; Yang 2001). Recent simulations confirmed that LRTs for detecting positive selection are conservative (Anisimova *et al.* 2001). Those observations suggest that genes inferred by the LRT to undergo positive selection are most likely to be true cases of adaptation rather than an artifact of the method.

Furthermore, in most of the published studies, positively selected sites inferred by the Bayes approach were biologically meaningful and/or clustered in a 3D structure of a protein while being dispersed in the primary sequence (*e.g.*, Bishop *et al.* 2000; Yang *et al.* 2000b; Peek *et al.* 2001; Yang and Swanson 2002).

However, patterns of genetic variability created by recombination can closely resemble the effects of molecular adaptation (*e.g.*, McVean 2001). With recombination, nucleotide sites in a sequence do not evolve along a single tree, but instead along a set of correlated trees (Hudson 1983). Recombination leads to apparent substitution rate heterogeneity (Worobey 2001). In phylogeny reconstruction, it is known to lead to star-like phylogenies and biases in tests of the molecular clock (Schierup and Hein 2000a,b). Current codon models of heterogeneous ω-ratios among sites assume no recombination, raising concerns about the possibility that the LRT can mistakenly interpret the effects of recombination as evidence for positive selection. Our previous simulation examined the accuracy of the LRT in nonrecombinant data sets (Anisimova *et al.* 2001).

In this article, we use computer simulation to investigate whether the LRT can lead to false detection of positive selection in the presence of recombination. We envisage that the problem mainly concerns viral genes, where sequence divergence is high and recombination may be frequent. Although recombination also occurs in population samples from other species such as animals and plants, the sequence divergence is in general too low for phylogeny-based likelihood methods to be useful (Anisimova *et al.* 2001). With this consideration of sequence divergence in mind, we simulate sequences using codon frequencies and parameter estimates obtained from a data set of the hepatitis D virus (HDV) antigen genes, in which both recombination and positive selection were reported (Wu *et al.* 1999).

MATERIALS AND METHODS

Likelihood-ratio tests: In this article we test the accuracy of likelihood-ratio tests for detecting positive selection at amino acid sites in the presence of recombination. We consider the following models of variable selective pressures among sites: M0 (one ratio), M1 (neutral), M2 (selection), M3 (discrete), M7 (β), and M8 (β and ω; Yang *et al.* 2000a). See Table 1. The simplest model, M0, assumes one ω-ratio for all sites. Model M1 allows two site classes, conserved sites with $\omega_0 = 0$ and neutrally evolving sites with $\omega_1 = 1$. Model M7 allows several site classes with ω-ratios drawn from the β distribution $B(p, q)$ and, hence, limited between 0 and 1. The three models, M0, M1, and M7, are taken as null hypotheses in the LRTs against their alternative models, M3, M2, and M8, respectively. Model M3 allows K discrete site classes with ω-ratios ω_0, ω_1, . . . , ω_{K-1} taken in proportions p_0, p_1, . . . , p_{K-1}. Here we use $K = 3$ as suggested by Yang *et al.* (2000a). Note that $K = 1$ in M0. Model M2 adds an extra class to M1 with an ω_2 estimated from data. Similarly, model M8 adds one discrete class to M7 with ω estimated from data. We consider three LRTs: (i) M0 (one ratio) *vs.* M3 (discrete), (ii) M1 (neutral) *vs.* M2 (selection), and (iii) M7 (β) *vs.* M8 (β and ω). We note that the M0-M3 comparison is really a test of variability of selective pressures among sites whereas the M1-M2 and M7-M8 comparisons are tests of positive selection.

We simulated data under the null model and analyzed them under both the null and the alternative models to calculate the LRT statistic $2\Delta\ell$ (twice the log-likelihood difference between the two models). For each data set, we reconstructed a neighbor-joining (NJ) tree using PAUP* (Swofford 2000). This tree was then used to perform codon-based likelihood analysis with the codeml program from the PAML package (Yang 2000). The statistic $2\Delta\ell$ was compared with the χ^2_ν distribution, with the degree of freedom ν equal to the difference in the number of free parameters between the two models; we used $\nu = 4$ for the M0-M3 comparison and $\nu = 2$ for the M1-M2 and M7-M8 comparisons. We note that strictly speaking, the asymptotic χ^2 approximation does not apply to those tests even without recombination and that our use of the χ^2 distribution makes the LRTs conservative (Anisimova *et al.* 2001). We counted the number of replicates in which the LRT was significant at $\alpha = 5$ and 1% (type I errors) and in which parameter estimates in the alternative models suggested positive selection; that is, ω_2 in M3 and M2 and ω in M8 were >1. Those were the cases in which positive selection was detected falsely.

TABLE 2

Number of replicates (out of 100) in the likelihood analysis comparing M0 (one ratio) and M3 (discrete)

Simulation parameters			Analysis under M0		Analysis under M3	
ω	θ_S	ρ	$\hat{\omega} > 1$	Mean of $\hat{\omega}$	At least one $\hat{\omega}_i > 1$	Significant at 5 (1)% and $\hat{\omega} > 1$
1	0.011	0	54	1.05	61	0 (0)
1	0.011	0.01	54	1.15	86	63 (61)
1	0.11	0	51	1.00	54	0 (0)
1	0.11	0.01	54	1.02	100	98 (94)
0.6	0.11	0	0	0.61	19	0 (0)
0.6	0.11	0.01	0	0.61	94	90 (89)
0.4	0.11	0	0	0.41	9	0 (0)
0.4	0.11	0.01	0	0.42	90	88 (88)

Data sets of 30 sequences were simulated under M0 and analyzed using both M0 and M3. Numbers of replicates significant at 1% are in parentheses.

Coalescent simulation with recombination: Sequence data were simulated by generating genealogies for sites in the sequence from a standard neutral coalescent model with recombination and then using them to "evolve" sequences according to a codon-substitution model. The genealogies are described by an *ancestral recombination graph*, generated by tracing the sample of DNA sequences backward in time while recording the coalescent and recombination events (*e.g.*, Hudson 1983; Griffiths and Marjoram 1996; Nielsen 2000). While selection clearly operates on a protein-coding gene, we ignore the effect of selection on the genealogy since currently no algorithm is available to simulate coalescent trees under both recombination and strong selection. The parameters involved in the coalescent simulation are $\theta = 2N\mu$ and $\rho = 2Nr$, where N is the effective population size, μ is the mutation (substitution) rate per codon site per generation, and r is the recombination rate per codon site per generation. In this article, we measure θ by the synonymous substitution rate and use the notation θ_S, calculated as d_S in Goldman and Yang (1994), with the expected coalescence time for a pair of sequences replacing branch length t. This is the expected number of synonymous substitutions per synonymous site between two sequences drawn at random from the population. Furthermore, since our mutation/substitution model is codon based, we allow recombination to occur between codons but not within a codon, so a site below refers to a codon (codon site).

The gene genealogy can be deduced at each site from the ancestral recombination graph. Evolution of sequences along the genealogy for each site (codon) can be simulated using standard methods (*e.g.*, Anisimova *et al.* 2001). In brief, a continuous-time Markov chain with rate matrix $Q = \{q_{ij}\}$ is superimposed along each lineage of the genealogy of a site. Multiple substitutions at the same site are thus allowed. The data were simulated using a C program written by R.N.

Values of parameters used in the simulation: All data sets were generated using codon frequencies and ML parameter estimates obtained from 33 geographically dispersed strains of a hepatitis D small antigen gene. The GenBank accession numbers of the hepatitis D antigen strains are AB015442, AB01543, AB015446, AB015447, AB037947–AB037949, AF018077, AF104263, AF104264, AF209859, AF309420, AJ309879, AJ309880, D01075, L22063, L22066, M28267, M58299, M58301, M58303, M58305, M58629, M84917, M92448, U19598, U25667, U81988, U81989, X04451, X63373, X77627, and X85253. The transition/transversion ratio was fixed at $\kappa = 3$.

We simulated data sets of 10 or 30 lineages under the null models M0 (one ratio), M1 (neutral), or M7 (β), and analyzed them using both the null models and the alternative models M3 (discrete), M2 (selection), and M8 (β and ω). We used two levels of sequence divergence at silent sites: $\theta_S = 0.011$ and 0.11. The strength of selection was varied by changing the ω-parameter. Under M0 (one ratio), we used $\omega = 0.4$, 0.6, and 1, while under M1, we assumed $\omega_0 = 0$ and $\omega_1 = 1$ in equal proportions. Under M7, the β-parameters were estimated from the HDV data: $p = 0.23$, $q = 0.41$. Reliable estimates of the scaled recombination rate ρ are unavailable for viral genes. As a result, we simulated data with different levels of recombination: $\rho = 0$, 0.0001, 0.0005, 0.001, 0.005, and 0.01, with most of the simulations done using only $\rho = 0$ (no recombination) and 0.01 (high level of recombination). Details of the parameter values used in the simulation are given in the results.

In many viral genes, both recombination and positive selection may be operating. We thus used simulation to examine the effect of recombination on identification of positively selected sites by the likelihood method. We simulated 30- and 10-taxa data sets using the alternative model M3 (discrete), assuming ~13.5% of sites under positive selection with $\omega_2 = 2.55$ and the remaining 65.8 and 20.6% of sites with $\omega_0 = 0.08$ and $\omega_1 = 0.61$, respectively, as estimated from the HDV data set. All replicates were moderately divergent ($\theta_S = 0.11$) and only two levels of recombination were used ($\rho = 0$ and $\rho = 0.01$). For 10-taxa data sets we varied the strength of positive selection while keeping other parameters the same: the ω_2 values were 2.55 and 6. Data were analyzed using alternative models, and sites inferred by the codeml program to be under selection were compared with the truly selected sites during the simulation. In all simulations the sequence length was 500 codons, while the number of replicates was 100.

RESULTS

Impact of recombination on the LRT: Table 2 shows the results for the likelihood analysis comparing M0 (one ratio) and M3 (discrete), performed on large data sets of 30 sequences. The data were simulated under M0. For nonrecombinant data ($\rho = 0$), the LRT did not reject the null model (M0) in any of the 100 replicates, regardless of the level of selective pressure (ω) or silent mutation rate (θ_S). The type I error rate was consistently

TABLE 3

Number of replicates (out of 100) in the likelihood analysis comparing M1 (neutral) and M2 (selection)

Simulation parameters		Analysis under M2		
θ_S	ρ	Significant at 5 (1)%	$\hat{\omega}_2 > 1$	Significant at 5 (1)% and $\hat{\omega}_2 > 1$
0.011	0	2 (0)	45	2 (0)
0.011	0.01	74 (61)	100	74 (61)
0.11	0	2 (2)	54	0 (0)
0.11	0.01	80 (71)	99	80 (71)

Data sets of 30 sequences were simulated under M1 (with $\omega_0 = 0$ and $\omega_1 = 1$ in equal proportions). Numbers of replicates significant at 1% are in parentheses.

lower than the significance level ($\alpha = 1$ or 5%). This result is consistent with the previous observation that the use of the χ^2 makes the LRT comparing M0 and M3 conservative (Anisimova *et al.* 2001).

However, when the data were simulated with recombination ($\rho = 0.01$), the LRT falsely rejected the null model M0 in many replicates. The type I error rate was much greater than the significance value for all parameter combinations and was higher for more divergent sequences (with larger θ_S) and larger ω. The error rate was as high as 98% when $\omega = 1$, $\theta_S = 0.11$, and $\rho = 0.01$ at the $\alpha = 5\%$ significance level. Examination of parameter estimates under M3 suggests that recombination increased the number of replicates in which at least one of the ML estimates of the ω-ratios under M3 was >1 and that, in most such replicates, the M0-M3 comparison was significant. Recombination also inflated the estimates of ω under model M0, but the effect was minor (Table 2). When $\rho = 0.01$, the recombination rate appears to be quite high, indicating an average of 32 recombination events in the history of a sample of 10 sequences.

Next we simulated data under the neutral model M1 to test whether the LRT comparing M1 and M2 was affected by recombination. The results are shown in Table 3. In the absence of recombination ($\rho = 0$), the LRT was conservative, with the type I error rate lower than the significance level. With recombination ($\rho = 0.01$), the type I error rate increased dramatically to 74% ($\theta_S = 0.011$) and 80% ($\theta_S = 0.11$) at the $\alpha = 5\%$ significance level. Recombination increased the number of replicates in which the ML estimate $\hat{\omega}_2$ in model M2 was >1 (Table 3).

Table 4 summarizes results obtained for different recombination rates when the LRT was used to compare models M0 (one ratio) and model M3 (discrete). The average number of recombination events observed in the simulation is shown for each recombination rate. When the recombination rate was low, with $\rho < 0.001$ or <2.7 recombination events in the history of a sample of 10 sequences, the LRT was conservative, with the type I error rate lower than the significance level (Table 4). When $\rho = 0.001$, the type I error rate was very slightly higher than the significance level α. Yet the percentage of replicates with falsely detected positive selection was approximately equal to the significance level α (Table 4). Increasing recombination rate further made the LRT highly inaccurate: positive selection was falsely detected in 23% (for $\rho = 0.005$) and 54% (for $\rho = 0.01$) of replicates at $\alpha = 5\%$.

Tables 2 and 4 also suggest that the type I error rate of the LRT is higher on big trees with 30 lineages than

TABLE 4

Number of replicates (out of 100) in the likelihood analysis comparing M0 (one ratio) and M3 (discrete)

ρ	Average no. of recombination events	Analysis under M3			Mean of $\hat{\omega}$ under M0
		Significant at 5 (1)%	At least one $\hat{\omega}_i > 1$	Significant at 5 (1)% and $\hat{\omega}_i > 1$	
0	0.00	0 (0)	14	0 (0)	0.4080
0.0001	0.23	0 (0)	17	0 (0)	0.4083
0.0005	1.27	1 (0)	21	0 (0)	0.4162
0.001	2.71	5 (2)	19	4 (2)	0.4120
0.005	15.50	36 (28)	46	23 (19)	0.4208
0.01	32.42	69 (58)	70	54 (46)	0.4224

Data sets of 10 sequences were simulated under M0 ($\omega = 0.4$, $\theta_S = 0.11$). For none of replicates was the estimate $\hat{\omega}$ under M0 > 1. Numbers of replicates significant at 1% are in parentheses.

TABLE 5

Number of replicates (out of 100) in the likelihood analysis comparing M7 (β) and M8 (β and ω)

	Analysis under M8		
ρ	Significant at 5 (1)%	$\hat{\omega} > 1$	Significant at 5 (1)% and $\hat{\omega} > 1$
0	8 (4)	47	4 (1)
0.001	7 (5)	49	3 (2)
0.01	21 (12)	81	20 (11)

Data sets of 10 sequences were simulated under M7 $B(p = 0.23, q = 0.41)$ with $\theta_S = 0.11$. Numbers of replicates significant at 1% are in parentheses.

on small trees with 10 lineages. For example, at the 5% significance level, the LRT comparing M0 with M3 failed in 88% of replicates for 30-lineage data sets and in 54% of replicates for 10-lineage data sets. Similarly, increasing the mutation/substitution rate leads to an increased false-positive rate in the LRT (Table 2).

Results obtained for the LRT comparing M7 (β) and M8 (β and ω) are summarized in Table 5. As those two models are computationally expensive, we used only three recombination levels: ρ = 0, 0.001, and 0.01. With no recombination (ρ = 0), the type I error rate was close to the significance level α (Table 5). A low recombination rate, with ρ = 0.001 or ~2.7 recombination events in a sample of 10 sequences, did not appear to make much difference in terms of the accuracy of the LRT (Table 5). Increasing ρ to 0.01, or ~32 recombination events in a sample of 10 sequences, caused the LRT to detect positive selection falsely in 20% of replicates at α = 5%. Although such an error rate is high, it is much lower than the error rate in the M0-M3 comparison for the same recombination rate (compare results for ρ = 0.01 in Tables 4 and 5).

The effect of incorrect phylogeny on the LRT: It is interesting to know why the LRT generates many false positives when the recombination rate is high. One possible reason is that the tree topology is incorrect for many sites since recombination causes different segments of the sequence to have different tree topologies. Schierup and Hein (2000a) suggested that even low levels of recombination can lead to biases in phylogenetic analyses. To examine the effect of assuming a "wrong" tree, we used a star topology to analyze the 10-taxa data sets generated in previous analyses (results not shown). The effect of using a star tree was profound: even for nonrecombinant data, the LRT falsely suggested positive selection in 96% of the replicates in the M0-M3 comparison and in 86% of the replicates in the M7-M8 comparison at the α = 5% significance level. Similarly high error rates for the LRT were observed when random tree topologies were used in the analysis (results not shown). Yet the ML estimates of ω under M0 (one ratio) were always very close to the true value, whichever tree was used (results not shown).

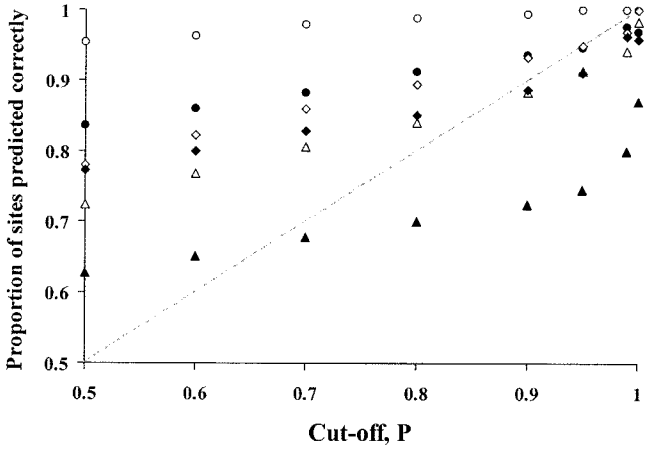

Figure 1.—Accuracy of Bayes' prediction of amino acid sites under positive selection, as measured by the proportion of identified sites that are truly under positive selection. The data were simulated under M3 (discrete) with 13.5% of sites under positive selection with $\omega_2 = 2.55$. The scaled recombination rate was ρ = 0 (no recombination) and 0.01 (high recombination). Each data set contained 30 sequences, simulated with $\theta_S = 0.11$ and was analyzed using models M2 (○, ρ = 0; ●, ρ = 0.01), M3 (△, ρ = 0; ▲, ρ = 0.01), and M8 (◇, ρ = 0; ◆, ρ = 0.01).

Bayes' prediction of sites under positive selection in the presence of recombination: In some genes, there is convincing evidence for both recombination and positive selection, and testing for the presence of sites under selection is not as important as identifying sites under selection. Thus we used simulation to evaluate the effect of recombination on the accuracy of Bayes' prediction of positive selection sites (Nielsen and Yang 1998; Yang *et al.* 2000a). Data sets with 30 lineages each were analyzed to identify sites under selection using three models: M2 (selection), M3 (discrete), and M8 (β and ω). We measured the accuracy of Bayes' prediction by the proportion of sites identified by the codeml program to be under selection that were truly under selection (Anisimova *et al.* 2002). Figure 1 shows the results of the analysis. When there was no recombination (ρ = 0), the accuracy of Bayes' site prediction was very high for M2 and M8, but slightly lower than predicted when data were analyzed using M3. For example, out of the sites predicted to be under positive selection at the 95% posterior probability cutoff, ~100 and 95% of sites were truly under positive selection when data were analyzed with M2 and M8, respectively, while the proportion was only 91% under M3. When recombination rate was high with ρ = 0.01, or an average of 46.7 recombination events in the history of a sample of 30 sequences, the accuracy of Bayes' site prediction decreased for all models (Figure 1). For M2, accuracy still remained high: at the 95% cutoff, ~95% of the inferred sites were correct. Accuracy of the analysis with M8 was only slightly lower

TABLE 6

Counts of replicates (out of 100) in the likelihood analysis comparing M0 (ω = 1 fixed) against M0 (ω estimated)

ρ	Significant at 5 (1)%	$\hat{\omega} > 1$	Significant at 5 (1)% and $\hat{\omega} > 1$	Mean of $\hat{\omega}$
NJ tree				
0	6 (1)	53	2 (1)	1.022
0.01	12 (4)	53	8 (2)	1.030
Star tree				
0	21 (13)	53	13 (9)	1.026
0.01	31 (20)	53	15 (12)	1.028

Data sets of 10 sequences were simulated under M0 with ω = 1 and θ_S = 0.11. Numbers of replicates significant at 1% are in parentheses.

than predicted: at the 95% cutoff, ~91% of the inferred sites were correct. However, when M3 was used for site identification, accuracy was much lower than predicted: at the 95% cutoff, only 75% of the inferred sites were predicted correctly. Such differences among models were also found by Anisimova *et al.* (2002) in simulations without recombination, where models M2 and M8 produced more accurate results than M3 produced, whichever model was used to simulate data.

Additionally, we investigated how the accuracy of Bayes' prediction of selected sites in the presence of recombination (ρ = 0.01) was affected by the number of sampled lineages and by the strength of positive selection. Models M2 and M3 were used to analyze data sets of 10 lineages. The accuracy of prediction (results not shown) was found to be very similar to that with 30 lineages (Figure 1). We also simulated data sets of 10 lineages using the same model M3 (discrete) but with the strength of positive selection increased from ω_2 = 2.55 to ω_2 = 6. We observed a substantial increase in both accuracy and power of Bayes' site prediction (results not shown).

DISCUSSION

Effect of recombination on the LRT: It is not surprising that the LRT becomes unreliable when recombination is frequent, since a basic assumption of the model is violated. Nevertheless, it is important to understand why the test fails at high recombination rates. As yet we do not have a good understanding of possible reasons for the failure of the LRT.

One possible reason is that recombination introduces variation in the tree length (the sum of branch lengths over the genealogy, measured in time) among sites, which introduces apparent variation in both synonymous and nonsynonymous substitution rates among sites. The codon models examined in this article account for variation only in nonsynonymous rates and assume a constant rate at synonymous sites. The synonymous rate is thus averaged over all sites. Sites with long tree lengths have more of both synonymous and nonsynonymous substitutions. At such sites the apparent nonsynonymous rate, even if lower than the synonymous rate at that site, can be higher than the average synonymous rate. Consequently, the method might incorrectly identify such sites as evolving under positive selection. If this interpretation is correct, accounting for variation in both synonymous and nonsynonymous rates will make the LRTs more robust to the presence of recombination.

This interpretation can be used to explain why the LRT of M7-M8 is much less affected than the LRTs of M0-M3 and M1-M2 by the presence of recombination. The M0-M3 comparison is a test for variability among sites. Variation in the tree length introduced by recombination can be seen as heterogeneity among sites in the gene tree and branch lengths. Thus, it can be expected that the test misinterprets such heterogeneity as variable ω-ratios. The case of the M1-M2 comparison is similar. Model M1 (neutral) accounts for only two site classes with ω_0 = 0 and ω_1 = 1 and is very unrealistic. As a result, model M2 (selection) misinterprets heterogeneity introduced by recombination as evidence for positive selection.

A second possible reason for the failure of the LRTs is that the tree topology, estimated by NJ for all sites in the sequence, becomes incorrect for many sites when recombination is frequent. It has been pointed out that recombination causes the estimated phylogeny to have long terminal branches resembling a star tree (Schierup and Hein 2000a; Worobey 2001). Consistent with this interpretation, we found that use of the star topology leads to many false positives in the LRT even when there is no recombination.

To explore this interpretation further, we examined another LRT, in which the null model was M0 with ω = 1 fixed, while the alternative model was M0 with ω estimated as a free parameter. The test statistic $2\Delta\ell$ was compared with the χ^2_1 distribution, using both the star tree and the NJ tree. For nonrecombinant data, the type I error rate at the α = 5% significance level was

2% when the NJ tree was used and 13% when the star tree was used (Table 6). The error rate when the NJ tree was used for recombinant data was 8% (Table 6). Those results appear to be consistent with our interpretation that recombination generates false positives partly because the reconstructed tree is wrong for some sites.

The effect of recombination on the LRT depends on when in the history of the sample recombination events have occurred. Recombinations in early internal branches are more disruptive of the genealogical tree and are expected to have greater effect than recent recombinations near the tips of the tree. If major recombination events can be identified in the sequence data, removing sequences involved in such events should increase the performance of the LRT. Factors that affect the shape of the genealogy and thus the distribution of recombination events on it are expected to affect the performance of the LRT as well. The present simulation is conducted under the neutral coalescent model, which generates trees with long internal branches. If the trees are more star-like, such as is the case when the population expands, recombination will have less effect than that found in this study. More importantly, both strong purifying and strong diversifying selection are known to operate in the viral genome, and genealogies under strong selection may be different from genealogies under neutrality. There is, as yet, no theory or algorithm for simulating coalescent trees under strong selection and recombination. Nevertheless, we expect that the effect of selection on the importance of recombination to the LRT, through its effect on the tree shape, is quantitative rather than qualitative, and our conclusions should remain valid. Studies showed that the shape of the genealogy was not seriously affected by weak-to-moderate purifying selection (Golding 1997; Przeworski *et al.* 1999; Slade 2000; Williamson and Orive 2002) or by background selection (Charlesworth *et al.* 1993, 1995; Hudson and Kaplan 1994, 1995).

Detecting positive selection in the presence of recombination: The simulations suggest that when the recombination rate is low, with fewer than about three recombination events in a sample of 10 sequences, the LRT is still accurate. However, much higher recombination rates cause the LRT to produce many false positives, sometimes as high as 100%. We found that Bayes' prediction of sites under positive selection is less affected by recombination. The reason seems to be that Bayes' identification of selected sites relies on reconstruction of the numbers of synonymous and nonsynonymous substitutions at individual sites, which is not very sensitive to the tree topology. In contrast to the LRT, increasing the number of lineages in the sample does not reduce the accuracy of Bayes' site prediction. Moreover, Bayes' site prediction becomes more accurate and powerful for higher levels of positive selection. We suggest that Bayes' site prediction may still be useful if positive selection is known to operate on the gene.

While the effect of recombination on the LRT depends on the recombination rate, reliable estimates of $\rho = 2Nr$ are unavailable for viral genes. The homoplasy index (Maynard and Smith 1998) and informative-sites index (Worobey 2001) are correlated with the recombination rate, but their exact relationships are unknown. More rigorous estimation methods are based on the coalescent model (Griffiths and Marjoram 1996; Kuhner *et al.* 2000; Nielsen 2000; Wall 2000; Fearnhead and Donnelly 2001; Hudson 2001). For human genes, most studies suggest the estimates of $\rho < 10^{-3}$/bp (*e.g.*, Hey and Wakeley 1997; Nielsen 2000). Such amounts of recombination have little effect on the LRT of positive selection; yet human population data typically lack variation so that the LRT is unlikely to detect an adaptive signal (Anisimova *et al.* 2001). For more divergent data, such as viral genes, estimation of recombination rates is much more problematic. Most methods are based on the neutral mutation model and do not account for variable selective pressures and thus do not account for variable substitution rates among sites in the gene or for regional positive correlation of substitution rates. As a result, they tend to mistake recurrent substitutions as evidence for recombination (McVean *et al.* 2002). Subsequently, the range of recombination rate estimates in viruses is very wide and there is no consensus on what rates might be reasonable. A number of studies discuss the possibility of recombination and positive selection both being present in data: *e.g.*, hepatitis D in Wu *et al.* (1999), foot-and-mouth virus in Haydon *et al.* (2001), fimbrial major subunit from *E. coli* in Peek *et al.* (2001), and apical membrane antigen 1 gene from malaria parasite *Plasmodium falciparum* in Polley and Conway (2001). While recombination is an evolutionary force maintaining genetic diversity, in some cases it can be seen as a strategy of evading the immune response, an alternative to diversification (*e.g.*, Burke 1997). Numerous reports of positive selection and recombination coexisting could also be an indication that current methods for detecting recombination and positive selection often confuse these two different forces, taking one for the other. McVean *et al.* (2002) extended the approximate-likelihood method of Hudson (2001) in an attempt to correct for a higher rate of recurrent mutations in viral and bacterial genes. Simulations showed that the method was more robust to misspecifications of the mutation model. However, it is unknown whether an excess of nonsynonymous substitutions at nonsynonymous sites causes an overestimation of recombination rate and whether the methods for detecting recombination are robust to variation of substitution rates among sites.

Clearly it is desirable to incorporate recombination into a coalescent codon-based model. The implementation would require the use of Markov chain Monte Carlo

approximation. Given the computational burdens of the current coalescent methods and the codon-based models, such methods are currently computationally intractable.

We thank two anonymous reviewers for improving the manuscript and Joseph P. Bielawski for discussions. This study was funded by grants from the Biotechnology and Biological Sciences Research Council to Z.Y., the Human Frontier Science Program to R.N. and Z.Y., and a National Science Foundation grant DEB-0089487 to R.N. M.A. is supported by a Medical Research Council studentship.

LITERATURE CITED

Anisimova, M., J. P. Bielawski and Z. Yang, 2001 Accuracy and power of the likelihood ratio test to detect 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.

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.

Burke, D. S., 1997 Recombination in HIV: an important viral evolutionary strategy. Emerg. Infect. Dis. **3:** 253–259.

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

Charlesworth, D., B. Charlesworth and M. T. Morgan, 1995 The pattern of neutral molecular variation under the background selection model. Genetics **141:** 1619–1632.

Fares, M. A., A. Moya, C. Escarmis, E. Baranowski, E. Domingo *et al.*, 2001 Evidence for positive selection in the capsid protein-coding region of the foot-and-mouth disease virus (FMDV) subjected to experimental passage regimens. Mol. Biol. Evol. **18:** 10–21.

Fearnhead, P., and P. Donnelly, 2001 Estimating recombination rates from population genetic data. Genetics **159:** 1299–1318.

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.

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

Golding, G. B., 1997 The effect of purifying selection on genealogies, pp. 271–285 in *Progress in Population Genetics and Human Evolution*, edited by P. Donnelly and S. Tavare. Springer-Verlag, New York.

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

Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. **3:** 479–502.

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.

Hey, J., and J. Wakeley, 1997 A coalescent estimator of the population recombination rate. Genetics **145:** 833–846.

Holmes, E. C., C. H. Woelk, R. Kassis and H. Bourhy, 2002 Genetic constraints and the adaptive evolution of rabies virus in nature. Virology **292:** 247–257.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805–1817.

Hudson, R. R., and N. L. Kaplan, 1994 Gene trees with background selection, pp. 140–153 in *Non-Neutral Evolution: Theories and Molecular Data*, edited by G. B. Golding. Chapman & Hall, New York.

Hudson, R. R., and N. L. Kaplan, 1995 Deleterious background selection with recombination. Genetics **141:** 1605–1617.

Kuhner, M. K., J. Yamato and J. Felsenstein, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156:** 1393–1401.

Maynard, S. J., and N. H. Smith, 1998 Detecting recombination from gene trees. Mol. Biol. Evol. **15:** 590–599.

McVean, G. A., 2001 What do patterns of genetic variability reveal about mitochondrial recombination? Heredity **87:** 613–620.

McVean, G., P. Awadalla and P. Fearnhead, 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160:** 1231–1241.

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154:** 931–942.

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.

Polley, S. D., and D. J. Conway, 2001 Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. Genetics **158:** 1505–1512.

Przeworski, M., B. Charlesworth and J. D. Wall, 1999 Genealogies and weak purifying selection. Mol. Biol. Evol. **16:** 246–252.

Schierup, M. H., and J. Hein, 2000a Consequences of recombination on traditional phylogenetic analysis. Genetics **156:** 879–891.

Schierup, M. H., and J. Hein, 2000b Recombination and the molecular clock. Mol. Biol. Evol. **17:** 1578–1579.

Slade, P. F., 2000 Simulation of selected genealogies. Theor. Popul. Biol. **57:** 35–49.

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

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.

Swofford, D. L., 2000 PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4.0b10. Sinauer Associates, Sunderland, MA.

Wall, J. D., 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.

Williamson, S., and M. E. Orive, 2002 The genealogy of a sequence subject to purifying selection at multiple sites. Mol. Biol. Evol. **19:** 1376–1384.

Worobey, M., 2001 A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria. Mol. Biol. Evol. **18:** 1425–1434.

Wu, J. C., T. Y. Chiang, W. K. Shiue, S. Y. Wang, I. J. Sheen *et al.*, 1999 Recombination of hepatitis D virus RNA sequences and its implications. Mol. Biol. Evol. **16:** 1622–1632.

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., 2000 Phylogenetic analysis by maximum likelihood (PAML), Version 3.0. University College, London.

Yang, Z., 2001 Maximum likelihood analysis of adaptive evolution in HIV-1 gp120 env gene. Pac. Symp. Biocomput., 226–237.

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

Yang, Z., and W. J. Swanson, 2002 Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol. Biol. Evol. **19:** 49–57.

Yang, Z., R. Nielsen, N. Goldman and A. M. Pedersen, 2000a Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:** 431–449.

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

Communicating editor: J. Hein