Reliabilities of identifying positive selection by the branch-site and the site-prediction methods

Masafumi Nozawa[a], Yoshiyuki Suzuki[a,b], and Masatoshi Nei[a,1]

[a]Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University; and [b]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics

Contributed by Masatoshi Nei, February 19, 2009 (sent for review January 9, 2009)

Natural selection operating in protein-coding genes is often studied by examining the ratio (ω) of the rates of nonsynonymous to synonymous nucleotide substitution. The branch-site method (BSM) based on a likelihood ratio test is one of such tests to detect positive selection for a predetermined branch of a phylogenetic tree. However, because the number of nucleotide substitutions involved is often very small, we conducted a computer simulation to examine the reliability of BSM in comparison with the small-sample method (SSM) based on Fisher's exact test. The results indicate that BSM often generates false positives compared with SSM when the number of nucleotide substitutions is ≈80 or smaller. Because the ω value is also used for predicting positively selected sites, we examined the reliabilities of the site-prediction methods, using nucleotide sequence data for the dim-light and color vision genes in vertebrates. The results showed that the site-prediction methods have a low probability of identifying functional changes of amino acids experimentally determined and often falsely identify other sites where amino acid substitutions are unlikely to be important. This low rate of predictability occurs because most of the current statistical methods are designed to identify codon sites with high ω values, which may not have anything to do with functional changes. The codon sites showing functional changes generally do not show a high ω value. To understand adaptive evolution, some form of experimental confirmation is necessary.

branch-site method | small-sample method

In the current statistical methods of inferring positive selection using the ω value, it is assumed that $\omega > 1$, $\omega = 1$, and $\omega < 1$ represent positive, neutral, and negative selection, respectively (1). One of the statistical methods using this approach is the branch-site method (BSM) (2, 3). In this method, the branches of a phylogenetic tree are divided into a predetermined (foreground) branch and other (background) branches and codon sites are grouped into a few classes with different ω values (see *Methods*). The log likelihood ($\ln L$) for the selection model used (modified model A) is then compared with that for the null model of no positive selection ($\omega \leq 1$), and the likelihood ratio test (LRT) is conducted to determine whether positive selection is operating in the foreground branch. This method has been widely used (e.g., 4–7), and one of the recent applications is Bakewell et al.'s (5) large-scale analysis of orthologous gene trios from humans, chimpanzees, and macaques. In this case, however, the numbers of synonymous (c_S) and nonsynonymous (c_N) substitutions per gene per branch were so small that the applicability of the large-sample theory of LRT is questionable.

Another test that is applicable for this type of datasets is the small-sample method (SSM) using Fisher's exact test (8). In this method the ancestral nucleotide sequence at each interior node is inferred by the parsimony method, and c_S and c_N for the branch to be tested are counted by comparing the sequences at the 2 terminal nodes of the branch. Positive selection is inferred when the c_N/c_S ratio is significantly higher than the ratio under the assumption of no selection. When c_S and c_N are small, the probability of occurrence of 2 or more substitutions at the same nucleotide site is negligibly small, and therefore parsimony estimates of c_S and c_N must be quite accurate. This is true even if the substitution rate varies with codon site to some extent. SSM should then be applicable for the primate dataset, and the results can be compared with those of BSM.

The ω value has also been used for predicting the positively selected codon sites in protein-coding genes (9–12). However, simulation studies showed that the Bayesian methods for predicting such sites often give false positives (13, 14). In fact, Yokoyama et al.'s (15) experimental study showed that these methods are not useful for identifying adaptive sites. These authors engineered the ancestral proteins of the dim-light vision opsins (RH1) from vertebrates and experimentally determined the critical amino acid substitutions that affect the maximum absorption wavelength (λ_{max}) of the opsin (rhodopsin) encoded. Because the spectral tuning of λ_{max} and the environmental condition of species were well correlated, these amino acid changes were considered to be adaptive. However, the Bayesian methods could not identify any of these critical sites. Because the critical amino acid changes affecting λ_{max} have also been identified in color vision genes (ref. 16 for review), we can extend this type of analysis to these genes as well.

In this article, we first examine the reliability of BSM in comparison with SSM by using a computer simulation. We are particularly interested in evaluating the false-positive rates of BSM and clarifying their causes. We then study the reliabilities of the Bayesian and other statistical methods for detecting positively selected sites by using real sequence data.

Results

Computer Simulation Mimicking the Primate Data. Our computer simulation for studying the reliability of BSM was done by mimicking Bakewell et al.'s (5) analysis of genes from the human-chimpanzee-macaque trios (Fig. 1). These authors considered the human or chimpanzee lineage as the foreground branch and the remaining lineages as the background. They examined ≈14,000 (actually 13,888) orthologous genes with an average of ≈450 (actually 432) codons. The average numbers of synonymous substitutions per synonymous site (b_S) for the human, chimpanzee, and macaque lineages were ≈0.006, ≈0.006, and ≈0.06, respectively. [In the following, we use the notation b_S = (0.006, 0.006, 0.06) for this case.] The average ω over all codon sites in these lineages was ≈0.25 (17). The transition/transversion rate ratio (κ) was ≈4 (18). On the basis of this information, we generated 14,000 sets of the human-chimpanzee-macaque trio sequences by a computer simulation (Fig. 1; see *Methods* for details). Because the ω value used was 0.25 for all sites, there were no sites under positive selection. Therefore, any site with an estimate ($\hat{\omega}$) of $\omega > 1$ must be caused by sampling or estimation errors. When we applied BSM for the 14,000 sets of genes considering the human lineage as the foreground branch, we obtained 32 genes showing positive selection at the 5% significance level ($\alpha = 0.05$) by using the computer program PAML 4 (19)

Author contributions: M. Nozawa and M. Nei designed research; M. Nozawa and Y.S. analyzed data; and M. Nozawa, Y.S., and M. Nei wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: nxm2@psu.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0901855106/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.0901855106

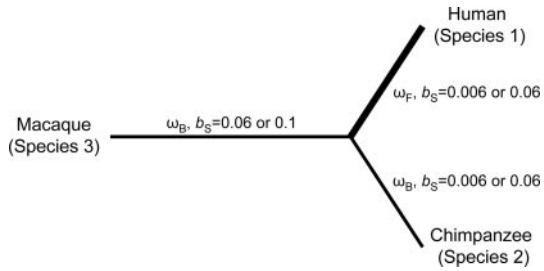

Fig. 1. Phylogenetic tree showing the simulation scheme. The foreground branch is shown by a bold line. In all simulations, κ is 4 and the number of codon sites (n) is 450. Not to scale.

(Table 1). (The results of BSM were obtained by PAML 4 unless otherwise stated.) By contrast, SSM based on Fisher's exact test showed no genes suggesting positive selection at the same significance level.

One might wonder why SSM did not detect any positive selection. The reason is that c_S and c_N were both too small to give any statistical significance. In the case of SSM, positive selection is suggested only when c_N is significantly greater than c_S, and for this to happen c_N must be 9 or greater even when $c_S = 0$ (Table S1). In practice, c_N was always equal to or smaller than 7 except for 2 cases, in which the c_N/c_S was 8/2 and 9/3. This indicates that the statistical information of the dataset used is not enough to get a significant result and that all of the 32 cases obtained by BSM are false positives.

Table 1. False-positive cases ($P < 0.05$) obtained by BSM in a computer simulation with $n = 450$, $\kappa = 4$, $\omega_F = \omega_B = 0.25$, and $b_S = (0.006, 0.006, 0.06)$

	BSM		SSM			BSM (HyPhy)	
Case	P	$\hat{\omega}_2$	P	c_S (true)*	c_N (true)	P	$\hat{\omega}_2$
1	0.036	999	0.11	0 (0)	6 (6)	0.039	18
2	0.004	999	0.92	2 (2)	2 (2)	0.005	943
3	0.016	593	0.92	2 (3)	2 (1)	0.016	582
4	0.011	386	0.54	1.5 (1)	3.5 (4)	0.014	383
5	0.013	464	0.53	1.8 (2)	3.5 (4)	0.018	482
6	0.030	999	0.99	4.5 (5)	1.5 (1)	**0.052**	**4,029**
7	0.006	472	0.92	2.5 (2)	1.5 (1)	0.013	492
8	0.007	999	0.85	2 (3)	3 (3)	**1.000**	**1.4**
9	0.001	999	0.91	2.5 (3)	1.5 (1)	0.001	1,138
10	0.028	102	0.74	2.5 (3)	3.5 (3)	0.037	103
11	0.006	242	0.97	3 (3)	2 (2)	0.010	244
12	0.006	271	0.83	2 (2)	3 (3)	0.007	280
13	0.005	349	0.92	2 (2)	2 (2)	0.008	359
14	0.010	195	0.99	5 (5)	3 (3)	0.020	209
15	0.011	102	0.64	2 (2)	5 (4)	0.014	105
16	0.003	548	0.66	1 (1)	3 (3)	0.003	492
17	0.003	999	0.64	2.5 (3)	4.5 (4)	0.001	10,000
18	0.031	106	0.90	4.5 (4)	4.5 (5)	0.031	103
19	0.021	152	0.74	2.5 (3)	3.5 (3)	0.027	156
20	0.002	888	0.78	1 (1)	2 (2)	0.003	860
21	0.002	999	0.79	1.5 (2)	1.5 (1)	0.003	1169
22	0.009	339	0.53	1.5 (2)	3.5 (3)	0.013	345
23	0.027	245	0.93	3 (4)	3 (2)	0.044	326
24	0.034	93	0.97	5 (5)	4 (4)	0.034	93
25	0.017	74	0.94	4 (4)	4 (4)	0.026	76
26	0.027	154	0.55	1.3 (1)	6.3 (7)	0.026	141
27	0.002	869	0.67	1 (1)	3 (3)	0.002	893
28	0.007	999	0.99	4 (4)	2 (2)	**1.000**	**1.0**
29	0.001	999	0.24	0 (0)	4 (5)	0.001	1,750
30	0.007	438	0.66	1.5 (2)	2.5 (2)	0.007	417
31	0.024	159	0.88	3.5 (4)	3.5 (3)	0.032	164
32	0.001	999	0.35	0 (0)	3 (3)	0.001	2827

BSM, branch-site method. SSM, small-sample method. In BSM, π was assumed to be equal for all 61 codons. The results for BSM and BSM (HyPhy) were obtained by the programs PAML 4 (19) and HyPhy (20), respectively. Nonsignificant cases by HyPhy are shown in bold. *, Estimated (true) numbers of substitutions in the human lineage in Fig. 1.

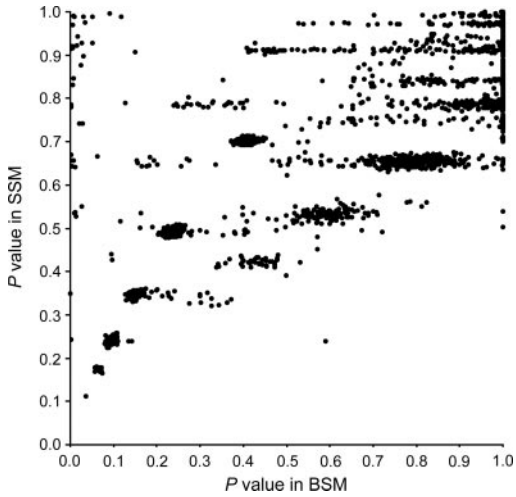

Fig. 2. Relationship between the P values for BSM and SSM. $n = 450$, $\kappa = 4$, $\omega_F = \omega_B = 0.25$, and 14,000 replications.

In the present simulation, we used the parsimony method for estimating c_S and c_N. However, because we recorded all mutations in the evolutionary process using a discrete-time model, we can use the true values of c_S and c_N for SSM. Table 1 shows that the parsimony estimates are often decimal, because there are 2 or more equally parsimonious pathways when 2 or more nucleotide differences exist between the 2 codons compared (21). In the computer simulation, we can identify which pathway was used so that the true numbers of substitutions are always integer. When these true numbers of c_S and c_N were used, virtually no changes in P values (type I error rates) occurred in SSM, and none of the genes showed positive selection at the 5% significance level. Some authors (22) have been critical of parsimony estimates of c_S and c_N and consequently of the methods based on parsimony estimates. In the present case, however, SSM is clearly more reliable than BSM.

Intuitively, one might expect that the P value for BSM becomes low when the c_N/c_S ratio is high, but Table 1 shows that this is not necessarily the case. This result was obtained apparently because LRT is affected by sampling errors seriously when the number of nucleotide substitutions is small and the regularity conditions for the χ^2 approximation are not satisfied in this method (3). Fig. 2 shows that the P value for SSM was always >0.1, indicating that the small c_S and c_N values are not informative for generating a significant result in any replication.

Table 1 also shows the estimates ($\hat{\omega}_2$) of ω ($= \omega_2$) for the group of codons for which positive selection was inferred by BSM (see *Methods*). A surprising observation is that the $\hat{\omega}_2$ values in the false-positive cases are all >70 and some are as high as 999, which is the maximum value that is printed by PAML 4 (19). Analyzing all 14,000 cases, we found that the $\hat{\omega}_2$ value tends to be higher when P is small than when P is large (Fig. S1). The average $\hat{\omega}_2$ value for 14,000 cases was 56.6. These $\hat{\omega}_2$ values are obviously erroneous because the true value is 0.25 for all codons.

One might argue that there is no need to worry about this type of abnormal behaviors of LRT because the observed false-positive rate (32/14,000 = 0.23%) is lower than the expected rate (5%) in large-sample tests. In the present case, however, BSM produces significant results when these results are not supposed to be obtained theoretically. This indicates that there is a computational problem in BSM. In addition, the false-positive rate in BSM can be >5% even under the condition of $\omega \leq 1$, as will be shown below.

EVOLUTION

Table 2. Numbers (percent) of false positives obtained by BSM and SSM in a computer simulation with $n = 450$, $\kappa = 4$, and 1,000 replications

		BSM	SSM		
ω_F	ω_B	Number (%)	Number (%)	True no. (%)*	$D^\dagger$
		A. b_S = (0.006, 0.006, 0.06)			
0.25	0.25‡	32 (0.2)	0 (0.0)	0 (0.0)	0.078
	0.5	4 (0.4)	0 (0.0)	0 (0.0)	0.112
0.5	0.25	16 (1.6)	0 (0.0)	0 (0.0)	0.060
	0.5	10 (1.0)	0 (0.0)	0 (0.0)	0.074
	1	9 (0.9)	0 (0.0)	0 (0.0)	0.127
1	0.25	59 (5.9)	18 (1.8)	18 (1.8)	0.037
	0.5	52 (5.2)	8 (0.8)	9 (0.9)	0.054
	1	64 (6.4)	10 (1.0)	12 (1.2)	0.086
		B. b_S = (0.06, 0.06, 0.1)			
0.25	0.25	6 (0.6)	0 (0.0)	0 (0.0)	0.078
	0.5	6 (0.6)	0 (0.0)	0 (0.0)	0.089
0.5	0.25	7 (0.7)	0 (0.0)	0 (0.0)	0.068
	0.5	11 (1.1)	0 (0.0)	0 (0.0)	0.076
	1	7 (0.7)	0 (0.0)	0 (0.0)	0.091
1	0.25	73 (7.3)	44 (4.4)	47 (4.7)	0.074
	0.5	48 (4.8)	29 (2.9)	44 (4.4)	0.081
	1	69 (6.9)	36 (3.6)	31 (3.1)	0.079

*Number (percentage) of false positives when the true c_S and c_N were used for SSM.

†Average deviation of parsimony estimates of c_S and c_N from the true values was measured by

$$D = \frac{1}{T}\Sigma \frac{|c_S^e - c_S^t| + |c_N^e - c_N^t|}{c_S^t + c_N^t},$$

where T is the number of replications and superscripts e and t refer to the estimated and true numbers of substitutions, respectively. When $c_S^t + c_N^t = 0$, we used 0.5 for the denominator in the computation.

‡Results are based on 14,000 replications.

False-Positive Rates when ω Varies with Branch. So far we considered only the case where $\omega = 0.25$ for both foreground and background branches. In reality, ω may be different between the foreground (ω_F) and background (ω_B) branches because of changes in functional constraints of the gene or some other factors. We have therefore considered several cases of different ω_F and ω_B values (Table 2, A). In this simulation, we generated 1,000 sets of genes for each case. In the first case of Table 2, A, we assumed $\omega_F = 0.25$ and $\omega_B = 0.5$. All other parameters were the same as before. In this case, 4 of the 1,000 genes showed positive selection in BSM, but none was detected by SSM. When we assumed $\omega_F = 0.5$ and $\omega_B = 0.25$, 0.5, or 1.0, BSM falsely detected positive selection with appreciable frequencies (0.9–1.6%), but SSM showed none. When $\omega_F = 1$ and $\omega_B = 0.25$, 0.5, or 1.0, the false-positive rate of BSM was ≈6% irrespective of the ω_B value. This rate is slightly higher than the expected false-positive rate (5%) in large-sample tests. By contrast, the false-positive rates for SSM were still ≈1%. This low rate indicates that the number of nucleotide substitutions was still quite small to be used for detecting positive selection. Because SSM is based on Fisher's exact test and the errors introduced by parsimony estimation of nucleotide substitutions are minor, the P value for BSM is apparently inflated by sampling errors. To see the effect of gene size, a similar simulation was conducted by using 900 codons instead of 450 codons. However, the results obtained were essentially the same as those of Table 2, A.

In the above computer simulation, the number of nucleotide substitutions per site was small because the human-chimpanzee-macaque trio was considered. However, BSM is also used for a group of species, which are more genetically divergent. We therefore extended our simulation to the case of greater b_S values [b_S = (0.06, 0.06, 0.1)] for species 1, 2, and 3 (Fig. 1). Roughly speaking, $b_S = 0.06$ for the foreground branch corresponds to the divergence time of ≈60 million years (MY) if humans and chimpanzees diverged ≈6 MY ago, whereas the total divergence time for the 3 species is ≈80 MY [corresponding to $b_S = (0.06 + 0.1)/2 = 0.08$]. Therefore, these divergence times are similar to those of primates and placental mammals, respectively (23).

The results of this simulation are presented in Table 2, B. The false-positive rates for BSM were more or less the same as those in Table 2, A. These results indicate that a larger number of nucleotide substitutions (e.g., ≈80 substitutions when $\omega_F = 1$) do not improve the reliability of BSM. In the case of $\omega_F = 1$, however, SSM showed a higher false-positive rate when b_S is large than when it is small, as expected from the increased number of nucleotide substitutions. Yet, the false-positive rates were still <5% and lower than those by BSM. In addition, there was no significant case when $\omega_F = 0.25$ or 0.5. Note that the true values of c_S and c_N gave essentially the same results in SSM (Table 2), indicating that the parsimony estimates of c_S and c_N are quite accurate. This accuracy was also confirmed by the small values (mostly <10%) of the average deviation (D) of parsimony estimates from the true values (Table 2).

These results were obtained from the simulation based on a discrete-time model with a time unit of $b_S = 0.0005$ (see *Methods*). We also conducted a simulation with a continuous-time model, using the computer program "evolverNSbranches.exe" in PAML 4. The results of the simulation using this model were essentially the same as those for the discrete-time model (Tables S2 and S3 and Fig. S2). It should be noted that a similar computer simulation mimicking the sequence evolution of the human-chimpanzee-macaque trios was conducted by Bakewell et al. (5) and Suzuki (24). Suzuki examined the false-positive rates for the cases of $\omega_F = 1$ and $\omega_B = 0.25$ or 1 with 450 and 750 codons and obtained rates of 7–8% instead of the expected rate of 5% in large-sample tests. Bakewell et al.'s simulation for the cases of $\omega_F = 1$ with 400 and 1,000 codons also showed a false-positive rate of 6–8%. These results show excessively high false-positive rates although the rate for SSM was not computed in these studies.

Data Analysis for Evaluating the Accuracy of Site-Prediction Methods. In addition to BSM, the ω approach is used for detecting positively selected sites. For this purpose, the Bayesian [M8 (9), and REL (11)] and likelihood [FEL (11)] methods are commonly used. We therefore examined the reliabilities of these methods, using real data. The data used here were the dim-light and color vision genes [RH1, RH1-like (RH2), short wavelength-sensitive type 1 (SWS1), SWS type 2 (SWS2), and middle and long wavelength-sensitive (M/LWS) genes] in vertebrates. In these genes potentially adaptive amino acid substitutions that affect the optimal light sensitivity measured by λ_{max} have been experimentally identified (e.g., 15, 16, 25). Using this information, we compared statistically predicted sites of positive selection with experimentally determined adaptive sites.

The results are shown in Table 3. In RH1 genes, many sites were predicted by the Bayesian methods and one site by the likelihood method when squirrelfish species were used. However, none of these sites agreed with the experimentally determined adaptive sites. In addition, most of the predicted sites disappeared when all vertebrate species were used for the analysis, as reported in ref. 15. In RH2 genes, the Bayesian methods did not detect any sites, and one site detected by the likelihood method did not agree with the adaptive sites experimentally determined. In SWS1 and M/LWS genes a few adaptive sites were correctly identified by the statistical methods when closely related species were used. Yet, most of the adaptive sites could not be detected by these methods. In SWS2 genes none of the sites was predicted as positively selected. These results indicate that in most cases the current statistical methods for site-prediction with the ω value cannot detect the adaptive sites, and instead they often falsely identify other sites as positively selected.

A different method called the DEPS method (28) was recently

Table 3. Positively selected sites by the site-prediction methods and experimentally determined adaptive sites in dim-light and color vision genes in vertebrates

Gene	Species group*	Predicted sites				Adaptive sites†
		M8 (Bayes)	REL (Bayes)	FEL (ML)	DEPS	
RH1	Squirrelfish (11)	37, 50, 162, 213, 214, 217	37, 39, 50, 54, 57, 112, 116, 162, 209, 213, 214, 217, 255, 256, 304	214	None	83, 96, 102, 122, 183, 194, 195, 253, 261, 289, 292, 317
	Vertebrates (38)	None	None	54	49, ***83***, 84, 88, 97, 100, 112, 137, 157, 165, 173, 189, ***194***, ***195***, 210, 214, 218, 248, 255, 260, 270, 277, 278, ***292***, 299, 304, 308, 309, 315	
RH2	Fishes (9)	None	None	104	None	49, 52, 83, 86, 97, 122, 164, 207, 292
	Vertebrates (15)	None	None	None	13, 116	
SWS1	Mammals (6)‡	38, 50, ***52***, ***97***, 100, 119, 159, 224, 332, 333	38, 50, ***52***, ***97***, 100, 159, 224, 332, 333	None	None	46, 49, 52, 86, 90, 91, 93, 97, 109, 113, 114, 116, 118
	Vertebrates (21)	None	None	None	61, ***114***, 273, 298, 315, 318, 320	
SWS2	Fishes (7)	None	None	None	None	44, 46, 91, 94, 109, 116, 118, 122, 261, 265, 269, 292
	Vertebrates (14)	None	None	None	None	
M/LWS	Primates (14)	65, ***180***, 229, 233, 275, ***285***	***180***	***180***	229	180, 197, 277, 285, 308
	Vertebrates (40)	275	None	None	55, 62, 140, 166, ***180***, 229, 278, ***285***	

For M8 (9), we used the program codeml.exe in PAML 4 (19). For REL and FEL (11), we used DATAMONKEY (26). In the Bayesian methods, positive selection was inferred if the posterior probability of $\omega > 1$ was 0.95 or higher for a site. In M8, the posterior probability for each site was computed for three times with different initial ω values (0.5, 1.5, and 2.5), and the results where the highest ln*L* was obtained are presented. In this method, we considered two different approaches, the naive-empirical-Bayes and Bayes-empirical-Bayes methods (27). If one of the approaches gave statistical support for a site, the site was regarded as positively selected. In the likelihood method, positive selection was inferred if the *P* value of LRT was less than 0.05 for a site. In DEPS, HyPhy (20) was used with the JTT model as the baseline matrix. Directional positive selection was inferred when the Bayes Factor was equal to or larger than 100. Correctly predicted sites are shown in bold italic. The sites for which the experiments of λ_{max} have not yet been conducted are underlined (S. Yokoyama, personal communication). Accession numbers for the sequences used are presented in *SI Text*. Bayes, Bayesian method; ML, likelihood method.

*The number of genes (including duplicate genes) in the entire set of species is given in parentheses.

†The critical codon sites where the amino acid substitutions change λ_{max}. Data are from Yokoyama et al. (15) and Yokoyama (16).

‡The sequences from primates, rodents, and bovine were used for analysis. In this case, all sites disappeared when only the Bayes-empirical-Bayes approach was considered for M8.

developed for predicting directional amino acid substitutions that may have changed protein function. This method does not rely on the ω value but uses the general pattern (baseline) of amino acid substitutions such as the JTT matrix. If a particular type of amino acid substitution occurs more frequently than the baseline matrix of amino acid substitutions, the amino acid substitution is assumed to be adaptive. The predicted sites by DEPS are also shown in Table 3. In RH1 genes, DEPS predicted 29 sites and 4 of them agreed with the experimentally determined adaptive sites. However, when only squirrelfish species were used, all of these predicted sites disappeared. In SWS1 and M/LWS genes, a few sites were correctly predicted as in the case of the ω based methods. In RH2 and SWS2 genes, however, none of the predicted sites agreed with the adaptive sites experimentally determined. (See Table S4 for the results obtained when the general time-reversible protein model was used.) These results indicate that DEPS also does not work well in predicting adaptive sites.

Why Does the Statistical Inference of Positive Selection Fail? One obvious answer to this question is the effect of sampling errors (13, 14), but the major factor for the failure appears to be the inadequacy of the mathematical model of nucleotide or amino acid substitution used. In both Bayesian and likelihood methods, synonymous substitutions are assumed to be neutral in the codon substitution model and the rate of nonsynonymous substitution is ω times higher than the rate of synonymous substitution, ω being the same for all nonsynonymous substitutions occurring in the same codon. Furthermore, the current Bayesian and likelihood methods all attempt to identify codon sites where the $\hat{\omega}$ value is significantly >1 and these sites are regarded to be under positive selection. For this reason, the average $\hat{\omega}$ value for these predicted sites is >1 even when $\hat{\omega}$ was computed by the conservative Suzuki-Gojobori method (29) (Table 4).

However, the average $\hat{\omega}$ value for experimentally determined

Table 4. Average $\hat{\omega}$ values for positively selected sites by site-prediction methods and experimentally determined adaptive sites in dim-light and color vision genes in vertebrates

Gene	Predicted sites*			Adaptive sites†
	M8 (Bayes)	REL (Bayes)	FEL (ML)	
RH1	6.09‡	4.42‡	5.50	0.40
RH2	—	—	1.04§	0.45
SWS1	2.47‡	2.57	—	0.67
SWS2	—	—	—	0.31
M/LWS	5.61‡	4.38	4.38	0.55

The $\hat{\omega}$ value for each codon site was computed by using the Suzuki–Gojobori method with $\kappa = 4$ in ADAPTSITE (29). Because the synonymous substitution rate varies considerably with codon site and often becomes 0, the average synonymous substitution rate over all codon sites was used for computing the $\hat{\omega}$ value for each codon site.

*The sequences from closely related species were used for the computation. RH1, squirrelfish; RH2, fishes; SWS1, mammals; SWS2, fishes; M/LWS, primates.

†All vertebrate species examined were used for the computation.

‡The value is significantly >1 at 5% significance level by the bootstrap test (21).

§This value is not so high, because we used the average synonymous substitution rate over all codon sites. However, c_S and c_N for the predicted site (site 104) were 0 and 4, respectively.

EVOLUTION

adaptive sites was much lower than that for the sites statistically inferred in all genes examined and was always <1 (Table 4). Why did this happen? The answer is that the functional change of a protein often occurs by replacement of a specific amino acid by another specific amino acid at 1 or few codon positions (see ref. 30 for review). For example, the SWS1 gene appears to have encoded a violet-sensitive opsin in the ancestor of birds, but the opsin became sensitive to UV in zebra finch, budgerigar, and canary (25). This change was caused by a single amino acid change from serine to cysteine at position 90 in the ancestor of these birds, and other amino acid changes were unlikely to be important (16, 25). In this case, it is quite difficult to detect this site by the statistical methods because adaptive substitution occurs very rarely. In fact, none of the statistical methods predicted this site even when only bird sequences were used. By contrast, the codon sites where many amino acid substitutions occurred may be falsely predicted as positively selected sites because of a high ω value that is obtained by chance even if the substitutions were essentially neutral (31). Note that >90% of amino acid substitutions are conservative and do not change protein function appreciably (15, 30, 32). For this reason, it is not easy to predict the evolutionary changes of protein function statistically.

A similar problem occurs with the DEPS method as well. In this method the amino acid changes that occur more often than the baseline expected from a given substitution matrix are regarded as adaptive. However, there is no reason to believe that these changes are adaptive, if only specific amino acid substitutions that occur rarely are adaptive. Furthermore, the theoretical basis of this method is not well established, because their baseline substitution matrix is not neutral but includes adaptive and conservative substitutions. Note that the baseline matrix is usually constructed from empirical data including all kinds of amino acid substitutions.

Discussion

We have shown that BSM gives false prediction of positive selection when the number of nucleotide substitutions in the foreground branch is small. This is apparently caused by the inadequacy of the statistical model used in BSM. For example, in the case of b_S = (0.006, 0.006, 0.06), 450 codons (n = 450), and ω = 0.25, the number of substitutions ($c_S + c_N$) in the foreground branch was ≈4 on average, but we have to estimate the 6 parameters, p_0, p_1, p_{2a}, p_{2b}, ω_0, and ω_2 from this small number (see *Methods*). (The number of independent parameters is 4 instead of 6, because p_{2a} and p_{2b} are computed from p_0 and p_1.) Obviously, the number of substitutions is insufficient for obtaining reliable estimates of the parameters.

In fact, the estimates ($\hat{p}_0, \hat{p}_1, \hat{p}_{2a}, \hat{p}_{2b}$, $\hat{\omega}_0$, and $\hat{\omega}_2$) of the parameters varied widely among different replications. For example, the estimates of the parameters for randomly chosen 5 nonsignificant and 5 significant replications in the case of b_S = (0.006, 0.006, 0.06) and $\omega_F = \omega_B = 1$ are given in Table 5 (see also Table S5). Because p_1 represents the proportion of class 1 sites that are assumed to be under no selection, $\hat{p}_1$ should be close or equal to 1 at least in nonsignificant cases. However, 5 nonsignificant cases showed that $\hat{p}_1$ varies from 0 to 0.97. The $\hat{\omega}_2$ value also ranged from 1 to 50, although this value should be 1 theoretically because no selection was assumed. In significant cases, $\hat{p}_1$ was 0 except in case 2 and $\hat{\omega}_2$ varied from 156 to 999. These wild variations of parameter estimates were apparently generated by sampling errors and the lack of the regularity conditions for the χ^2 approximation in LRT mentioned earlier. Note that $\hat{p}_2$ (sum of the estimates of the proportions of site classes 2a and 2b, which are assumed to be under positive selection for the foreground branch) was 1 in many cases (389 of 1,000 replications). This is unreasonable because no positive selection was assumed. Therefore, the results of BSM applied to primate data are not really reliable.

This unreliability of parameter estimates obtained by BSM is also revealed by their sensitivity to differences in the computational procedure. Table 1 shows the P and $\hat{\omega}_2$ values obtained by the computer program HyPhy (20) and PAML 4 (19). The maximization procedures of likelihood in the 2 programs are somewhat different, and this difference alone gave very different conclusions about the statistical significance in cases 8 and 28. In these two cases, PAML 4 gave a small P value, whereas HyPhy gave P = 1. Very different results were also obtained by PAML 4 and HyPhy for the case of the continuous-time model (Table S2). Note also that if we estimate the frequency of each codon (π) from the data rather than by assuming π = 1/61, the results of BSM again changes considerably in both PAML 4 and HyPhy (Table S6).

Table 5. Estimates for the six parameters in the BSM analysis of 5 nonsignificant and significant cases when b_S = (0.006, 0.006, 0.06) and $\omega_F = \omega_B = 1$

Case	$\hat{p}_0$	$\hat{p}_1$	$\hat{p}_{2a}$	$\hat{p}_{2b}$	$\hat{\omega}_0$	$\hat{\omega}_2$
Nonsignificant cases						
1	0.51	0.00	0.49	0.00	0.84	1.00
2	0.00	0.00	0.99	0.01	1.00	1.32
3	0.02	0.97	0.00	0.01	1.00	1.00
4	0.00	0.00	0.01	0.99	0.07	3.22
5	0.45	0.53	0.01	0.01	0.24	50.5
Significant cases						
1	0.00	0.00	1.00	0.00	0.72	999
2	0.01	0.97	0.00	0.02	1.00	228
3	0.00	0.00	1.00	0.00	0.96	175
4	0.00	0.00	0.00	1.00	1.00	999
5	0.98	0.00	0.02	0.00	0.93	156

See Table S5 for the average and standard deviation of the parameters for all 1,000 cases.

Another indication of the difficulty of obtaining reliable likelihood estimates of parameters by BSM is the fact that when multiple nonsynonymous substitutions occur in a codon the gene is often identified as positively selected even if no positive selection actually operates in the gene (24). We call this the Suzuki effect, and this effect causes an erroneous identification of positive selection when closely related species are studied. The well-known recommendation of the use of multiple initial ω values in PAML 4 is also a clear indication of difficulties of obtaining maximum likelihood estimates.

However, a more serious problem is the inadequacy of the ω approach, as was shown with respect to the statistical prediction of positively selected sites. If the ω approach is not applicable, BSM or any other method using ω would give questionable results. We have also shown that the prediction of adaptive amino acid substitution by the DEPS method (28) often disagrees with the experimentally determined adaptive sites.

What should we do if we want to study the adaptive significance of amino acid substitutions? The best way would be to use site-directed mutagenesis or similar techniques and study the functional or fitness change due to a specific amino acid substitution experimentally (16, 33). For example, it was experimentally shown that a high-virulence strain of West Nile virus in American crows is caused by a single amino acid substitution from threonine to proline at position 249 of NS3 helicase (34). A study combining experimental and statistical methods also identified natural selection in the digestive RNase genes in leaf-eating monkeys (35). In some proteins, however, this type of experimental study may be difficult to conduct. In such cases statistical tests of selection may be useful if the study is done by considering biochemical data available. For example, major histocompatibility complex (MHC) loci are known to be extraordinarily polymorphic, but the cause of this polymorphism was not known until Hughes and Nei (1, 36) showed that the ω value was significantly >1 at the antigen binding region (ABR) of the MHC molecules and the ABR tends to include amino acid substitutions that cause charge changes of the molecules (37). From these observations, Hughes and Nei (1, 36) concluded that the MHC polymorphism must be caused by some kinds of balancing selection. In general, statistical methods in combination of biolog-

ical information may be useful for immune systems or antigenic genes.

In the above discussion we considered a model in which ω varies with codon site within a gene. Originally, however, ω was proposed to measure the direction and extent of selection operating for an entire gene. In this case ω is computed by using the average rates of synonymous and nonsynonymous substitutions for the entire nucleotide sequences of a gene (21). For this purpose, ω is still useful because the sampling error of this ω is generally small. The c_N/c_S value for the entire gene used in SSM is also useful for detecting positive selection when a new function of a gene evolves as a result of many nucleotide changes for the same direction [e.g., generating cationic proteins (38)].

The fitness of an individual is a complex character and is determined by a large number of genes particularly with respect to morphological characters. Therefore, even if some gene experiences a functional change, it may not necessarily affect the fitness of the individual. It is interesting to note that even the selective advantage of trichromatic color vision over the dichromatic vision has been disputed in New World monkeys (39, 40). It is important not to be overenthusiastic about statistical signatures of positive selection without biological confirmation.

Methods

Computer Simulation. In generating DNA sequences we used the discrete-time model to compute the true c_S and c_N values for each evolutionary lineage. The 3 nucleotide sequences in Fig. 1 were generated by using the codon substitution model (19). The equilibrium frequencies were assumed to be the same for all 61 sense codons ($\pi = 1/61$) and $\kappa = 4$. The evolutionary time unit as measured by b_S was 0.0005 in our simulation. To confirm the accuracy of our computation, we also generated sequences by using the continuous-time model ''evolverNS-branches.exe'' in PAML 4 (19).

Statistical Methods. After generating sequences, we conducted the BSM analysis, using the program codeml.exe in PAML 4. In this method, the branches of a tree are divided into the foreground and the background branches. All codon sites are categorized into classes 0, 1, 2a, and 2b with proportions of p_0, p_1, p_{2a}, and p_{2b}, respectively. In the modified model A (Table S7), negative selection is assumed to operate on both the foreground and background branches ($0 < \omega_0 < 1$) in class 0. In class 1, no selection is assumed to occur for both the foreground and background branches ($\omega_1 = 1$). In class 2a, it is assumed that positive selection operates on the foreground branch ($\omega_2 > 1$), whereas negative selection operates on the background branches ($\omega = \omega_0$). In class 2b, positive selection is assumed to operate on the foreground branch ($\omega = \omega_2$), whereas no selection is assumed for the background branches ($\omega = \omega_1 = 1$). The null model of no positive selection is the same as the selection model except that no selection is assumed on the foreground branch ($\omega_2 = 1$) in classes 2a and 2b. The ln*L*s for these models are computed, and positive selection is inferred for the foreground branch if the LRT is greater than $\chi^2_1 = 3.84$ (5% significance level) (3). In this study, we computed ln*L*s 3 times, using 3 different initial ω values (0.5, 1.5, and 2.5) as recommended. The highest ln*L* among 3 trials was used for the computation of LRT. For each replication, we used the estimates ($\hat{p}_0$, $\hat{p}_1$, $\hat{p}_{2a}$, $\hat{p}_{2b}$, $\hat{\omega}_0$, and $\hat{\omega}_2$) of the 6 parameters for the highest ln*L*. We also conducted the BSM analysis, using the program ''YangNielsenBranchSite2005.bf'' in HyPhy to compare the results with those by PAML 4.

For SSM, the ancestral nucleotide sequence of humans and chimpanzees (or species 1 and 2) (Fig. 1) was inferred by the parsimony method. The c_S and c_N values and the numbers of synonymous and nonsynonymous sites in the human (or species 1) lineage were then estimated by using the modified Nei-Gojobori method (38) with the ratio of the numbers of transitions to transversions (R) = 2 (κ = 4). The test of neutrality was conducted by using Fisher's exact test (see Table S1).

Real Data Analysis. We used the dim-light and color vision (RH1, RH2, SWS1, SWS2, and M/LWS) genes in vertebrates. We obtained information about the critical amino acid changes for λ_{max} from the previous studies (15 for RH1 genes, 16 for other genes). The nucleotide sequences of these genes were obtained from the GenBank (see *SI Text* for accession numbers). For RH1 genes, the sequences were provided by Shozo Yokoyama. Detailed procedures are presented in Tables 3 and 4.

ACKNOWLEDGMENTS. We thank Hiroshi Akashi, Sayaka Miura, Naoko Takezaki, and Koichiro Tamura for their help in our computer simulation and Saby Das, Hiroki Goto, Eddie Holmes, Austin Hughes, Sergei Kosakovsky Pond, Bing Li, Bruce Lindsay, Jongmin Nam, and Shozo Yokoyama for their comments on earlier versions of the manuscript. This work was supported by National Institutes of Health Grants GM020293 (to M. Nei) and KAKENHI 17770007 (to Y.S.).

1. Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335:167–170.
2. Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* 19:908–917.
3. Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* 22:2472–2479.
4. Arbiza L, Dopazo J, Dopazo H (2006) Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Comput Biol* 2:e38.
5. Bakewell MA, Shi P, Zhang J (2007) More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc Natl Acad Sci USA* 104:7489–7494.
6. Kosiol C, et al. (2008) Patterns of positive selection in six Mammalian genomes. *PLoS Genet* 4:e1000144.
7. Studer RA, Penel S, Duret L, Robinson-Rechavi M (2008) Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Res* 18:1393–1402.
8. Zhang J, Kumar S, Nei M (1997) Small-sample tests of episodic adaptive evolution: A case study of primate lysozymes. *Mol Biol Evol* 14:1335–1338.
9. Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.
10. Suzuki Y (2004) New methods for detecting positive selection at single amino acid sites. *J Mol Evol* 59:11–19.
11. Kosakovsky Pond SL, Frost SD (2005) Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol* 22:1208–1222.
12. Massingham T, Goldman N (2005) Detecting amino acid sites under positive selection and purifying selection. *Genetics* 169:1753–1762.
13. Suzuki Y, Nei M (2001) Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. *Mol Biol Evol* 18:2179–2185.
14. Suzuki Y, Nei M (2002) Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. *Mol Biol Evol* 19:1865–1869.
15. Yokoyama S, Tada T, Zhang H, Britt L (2008) Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. *Proc Natl Acad Sci USA* 105:13480–13485.
16. Yokoyama S (2008) Evolution of dim-light and color vision pigments. *Annu Rev Genomics Hum Genet* 9:259–282.
17. Gibbs RA, et al. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316:222–234.
18. Rosenberg MS, Subramanian S, Kumar S (2003) Patterns of transitional mutation biases within and among mammalian genomes. *Mol Biol Evol* 20:988–993.
19. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.
20. Kosakovsky Pond SL, Frost SD, Muse SV (2005) HyPhy: Hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.
21. Nei M, Kumar S (2000) *Molecular Evolution and Phylogenetics* (Oxford Univ Press, New York).
22. Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* 15:496–503.
23. Murphy WJ, Pringle TH, Crider TA, Springer MS, Miller W (2007) Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res* 17:413–421.
24. Suzuki Y (2008) False-positive results obtained from the branch-site test of positive selection. *Genes Genet Syst* 83:331–338.
25. Shi Y, Yokoyama S (2003) Molecular analysis of the evolutionary significance of ultraviolet vision in vertebrates. *Proc Natl Acad Sci USA* 100:8308–8313.
26. Kosakovsky Pond SL, Frost SD (2005) Datamonkey: Rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* 21:2531–2533.
27. Yang Z, Wong WS, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22:1107–1118.
28. Kosakovsky Pond SL, Poon AF, Leigh Brown AJ, Frost SD (2008) A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. *Mol Biol Evol* 25:1809–1824.
29. Suzuki Y, Gojobori T, Nei M (2001) ADAPTSITE: Detecting natural selection at single amino acid sites. *Bioinformatics* 17:660–661.
30. Nei M (2005) Selectionism and neutralism in molecular evolution. *Mol Biol Evol* 22:2318–2342.
31. Hughes AL, Friedman R (2008) Codon-based tests of positive selection, branch lengths, and the evolution of mammalian immune system genes. *Immunogenetics* 60:495–506.
32. Perutz MF (1983) Species adaptation in a protein molecule. *Mol Biol Evol* 1:1–28.
33. Jermann TM, Opitz JG, Stackhouse J, Benner SA (1995) Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* 374:57–59.
34. Brault AC, et al. (2007) A single positively selected West Nile viral mutation confers increased virogenesis in American crows. *Nat Genet* 39:1162–1166.
35. Zhang J (2006) Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nat Genet* 38:819–823.
36. Hughes AL, Nei M (1989) Nucleotide substitution at major histocompatibility complex class II loci: Evidence for overdominant selection. *Proc Natl Acad Sci USA* 86:958–962.
37. Hughes AL, Ota T, Nei M (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.
38. Zhang J, Rosenberg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci USA* 95:3708–3713.
39. Regan BC, et al. (2001) Fruits, foliage and the evolution of primate colour vision. *Philos Trans R Soc Lond B Biol Sci* 356:229–283.
40. Hiramatsu C, et al. (2008) Importance of achromatic contrast in short-range fruit foraging of primates. *PLoS ONE* 3:e3356.