Genes Genet. Syst. (2008) **83**, p. 331–338

False-positive results obtained from the branch-site test of positive selection

Yoshiyuki Suzuki[1,2]*

[1]*Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University*
[2]*Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics*

(Received 8 April 2008, accepted 5 July 2008)

Natural selection operating at the amino acid sequence level can be detected by comparing the rates of synonymous (r_S) and nonsynonymous (r_N) nucleotide substitutions, where r_N/r_S (ω) > 1 and ω < 1 suggest positive and negative selection, respectively. The branch-site test has been developed for detecting positive selection operating at a group of amino acid sites for a pre-specified (foreground) branch of a phylogenetic tree by taking into account the heterogeneity of ω among sites and branches. Here the performance of the branch-site test was examined by computer simulation, with special reference to the false-positive rate when the divergence of the sequences analyzed was small. The false-positive rate was found to inflate when the assumptions made on the ω values for the foreground and other (background) branches in the branch-site test were violated. In addition, under a similar condition, false-positive results were often obtained even when Bonferroni correction was conducted and the false-discovery rate was controlled in a large-scale analysis. False-positive results were also obtained even when the number of nonsynonymous substitutions for the foreground branch was smaller than the minimum value required for detecting positive selection. The existence of a codon site with a possibility of occurrence of multiple nonsynonymous substitutions for the foreground branch often caused the branch-site test to falsely identify positive selection. In the re-analysis of orthologous trios of protein-coding genes from humans, chimpanzees, and macaques, most of the genes previously identified to be positively selected for the human or chimpanzee branch by the branch-site test contained such a codon site, suggesting a possibility that a significant fraction of these genes are false-positives.

Key words: branch-site test, positive selection, false-positive, computer simulation, real data analysis

INTRODUCTION

Natural selection operating at the amino acid sequence level can be detected by comparing the rates of synonymous (r_S) and nonsynonymous (r_N) nucleotide substitutions. Under the assumption that the synonymous substitution is selectively nearly neutral, r_N/r_S (ω) > 1 and ω < 1 suggest positive and negative selection, respectively (Hughes and Nei, 1988). In a protein, the biological function and the physicochemical environment not only vary at different amino acid sites but also change with time during evolution. Therefore, the direction and magnitude of natural selection may vary at different amino acid sites as well as for different branches of the phylogenetic tree under investigation.

The branch-site test has been developed for detecting positive selection operating at a group of amino acid sites for a pre-specified (foreground) branch of a phylogenetic tree by taking into account the heterogeneity of ω among sites and branches (Yang and Nielsen, 2002; Zhang et al., 2005). Multiple nucleotide sequences of protein-coding genes are compared under the assumption that their phylogenetic relationship is known. In the modified model A, codon sites are assumed to be classified into site classes 0, 1, 2a, and 2b, which exist with proportions p_0, p_1, p_{2a} [$= (1 - p_0 - p_1)p_0/(p_0 + p_1)$], and p_{2b} [$= (1 - p_0 - p_1)p_1/(p_0 + p_1)$], respectively. In site class 0, negative selection is assumed to operate for both the foreground and other (background) branches ($0 < \omega_0 < 1$). In site class 1, no

Edited by Hidenori Tachida
* Corresponding author. E-mail: yossuzuk@lab.nig.ac.jp

selection is assumed to operate for either the foreground or background branches ($\omega = 1$). In site class 2a, positive selection is assumed to operate for the foreground branch ($\omega_2 > 1$), whereas negative selection is assumed for the background branches ($\omega = \omega_0$). In site class 2b, positive selection is assumed to operate for the foreground branch ($\omega = \omega_2$), whereas no selection is assumed for the background branches ($\omega = 1$). The null model is the same as this model, except that no selection is assumed to operate for the foreground branch ($\omega = 1$) in site classes 2a and 2b. The likelihood function is formulated based on the codon substitution model (Goldman and Yang, 1994; Muse and Gaut, 1994), and free parameters are estimated by the maximum likelihood method. Positive selection is inferred to have operated for the foreground branch if the log-likelihood ($\ln L$) of the modified model A is greater than that of the null model at the 5% significance level in the likelihood-ratio test (LRT), which is conducted under the assumption that twice the difference in $\ln L$ ($2\Delta\ln L$) is asymptotically distributed as an equal mixture of a point mass on 0 and χ_1^2 However, since the test may be biased because of violation of the assumptions made for the model and a small sample size, it is recommended to use χ_1^2 to make the test conservative.

The branch-site test has been widely used. One of the recent applications of this test is Bakewell et al.'s (2007) large-scale analysis of 13,888 orthologous gene trios from humans, chimpanzees, and macaques, where positive selection was detected for 154 and 233 genes for the human and chimpanzee branches, respectively. Computer simulation mimicking this analysis was conducted, and it was concluded that the test was reliable (Bakewell et al., 2007). However, because the simulation was conducted for a limited number of conditions, the reliability of their conclusion remains unclear.

The purpose of the present study was to examine the performance of the branch-site test by computer simulation, with special reference to the false-positive rate when the divergence of the sequences analyzed was small. The significance of false-positive results in the real data analysis is discussed.

MATERIALS AND METHODS

Computer simulation Computer simulation was designed to mimic the analysis of orthologous trios of protein-coding genes from humans, chimpanzees, and macaques. The parameter values used in the simulation were obtained from the real data analysis. The analysis of 13,888 genes from these species showed that the average number of codon sites (n) in a gene was 432 (Bakewell et al., 2007). The average number of synonymous substitutions per synonymous site (b_S) for the human, chimpanzee, and macaque branches of the phylogenetic tree were approximately 0.006, 0.006, and 0.058, respectively. The average ω over all codon sites was 0.259, 0.245, and 0.226 for the human, chimpanzee, and macaque branches, respectively, and the average ω over all codon sites for all branches was 0.247 (Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007). The transition/transversion rate ratio (κ) was approximately 4 (Rosenberg et al., 2003; Jiang and Zhao, 2006).

On the basis of this information, three sequences with $n = 450$ were generated according to the phylogenetic tree as shown in Fig. 1 (Suzuki and Gojobori, 1999). b_S's were set to be 0.006, 0.006, and 0.06 for the human, chimpanzee, and macaque branches, respectively. It was assumed that the equilibrium frequencies for 61 sense codons were the same (1/61) and that $\kappa = 4$. ω was assumed to be 0.25 at all codon sites for both the foreground (ω_F) and background (ω_B) branches. In addition, since the parameter values may be heterogeneous among genes in the real data analysis, the simulation was also conducted under the assumptions that $n = 750$, $\kappa = 2$, $\omega_F = 1$, and $\omega_B = 0$, 1, and 5. In the present study, only one site class was assumed for generating sequences in the computer simulation, which was much simpler than in Bakewell et al. (2007), where up to 10 site classes were assumed. Although the assumption that $\omega_B = 5$ at all codon sites may be unrealistic, this was useful for clarifying the statistical properties of the branch-site test.

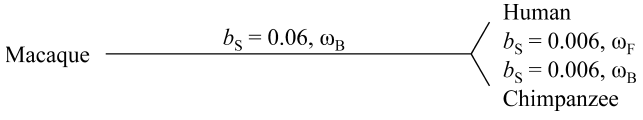

Fig. 1. Phylogenetic tree used for generating three sequences in the computer simulation.

The sequences generated were analyzed by the branch-site test using the computer program PAML (version 3.15) (Yang, 2007). The human branch was assumed as the foreground branch, but this was the same as assuming the chimpanzee branch as the foreground branch, because the lengths of the human and chimpanzee branches were the same (Fig. 1). The initial value of ω in the estimation of ω_0 and ω_2 was 0.4. The numbers of synonymous (c_S) and nonsynonymous (c_N) substitutions that have occurred for the human branch were estimated by comparing the ancestral nucleotide sequence at the interior node, which was inferred using the maximum parsimony (MP) method, with the human sequence. The true values of c_S and c_N were approximately obtained by comparing the true sequence at the interior node, which was available during generation of simulated sequences, with the human sequence. The entire process was repeated 14,000 times. Bonferroni correction for multiple testing was applied to the test. In addition, the false-

discovery rate (FDR) was also controlled to be 5% by using QVALUE (Storey, 2002).

Real data analysis The 13,888 genes from humans, chimpanzees, and macaques that were studied by Bakewell et al. (2007) were re-analyzed to examine the nucleotide substitutions that have occurred for the human and chimpanzee branches of the phylogenetic tree. The MP method was used to infer the ancestral nucleotide sequence at the interior node, and the nucleotide substitutions that have occurred for the human and chimpanzee branches were inferred by comparing the ancestral sequence with the human and chimpanzee sequences, respectively.

Minimum value of c_N required for detecting positive selection When c_S and c_N as well as the numbers of synonymous (s_S) and nonsynonymous (s_N) sites are given, positive selection can be detected by computing the exact P value for obtaining the observed or the greater value for c_N under the assumption that c_S and c_N follow a binomial distribution with the probabilities for occurrence of synonymous and nonsynonymous substitutions given by $s_S/(s_S + s_N)$ and $s_N/(s_S + s_N)$, respectively (one-tailed test). Therefore, the minimum value of c_N required for detecting positive selection at the 5% significance level is positively correlated with $s_N/(s_S + s_N)$ and c_S. s_S and s_N are different among 61 sense codons. However, when $\kappa = 4$, the smallest value of s_N is 1.33 ($s_S = 1.67$) for codons CTA and CTG, indicating that $s_N/(s_S + s_N)$ cannot be smaller than 0.44 for any sequence. Therefore, the minimum value of c_N required for detecting positive selection can be obtained as the value required under the assumptions that $s_N/(s_S + s_N) = 0.44$ and $c_S = 0$, and the value is 4. It should be noted that the minimum value is negatively correlated with κ. However, the value is 4 within the range of $3 \le \kappa \le 17$, which is likely to contain the true value of κ for humans, chimpanzees, and macaques (Rosenberg et al., 2003; Jiang and Zhao, 2006).

RESULTS

False-positive rates in the branch-site test Computer simulation was designed to mimic the analysis of 14,000 orthologous trios of protein-coding genes from humans, chimpanzees, and macaques (Fig. 1), by assuming various values for n, κ, ω_F, and ω_B. When $n = 450$, $\kappa = 4$, $\omega_F = 0.25$, and $\omega_B = 0.25$, positive selection was detected for the human branch in 99 cases (replications) by the branch-site test (Table 1). c_S and c_N for the human branch as inferred by the MP method for the 99 cases were generally small (≤ 6 for both c_S and c_N) and virtually identical with the true values, suggesting that the former values were reliable (Table 2). It should be noted that, when $\kappa = 4$, it appears to be impossible for a statistical test to detect positive selection if $c_N < 4$. However, among the 99 cases, inferred and true values of c_N were < 4 for 67 and 65 cases, respectively. These results suggest that most of the cases for positive selection identified by the branch-site test were errors, although the false-positive rate (0.71%) was < 5% in this case.

The false-positive rate did not change to any large extent when n and κ were assumed to be 750 and 2, respectively (Table 1). However, the rate varied according to the ω_B value. In particular, among the ω_B values (0, 0.25, 1, and 5) used in the simulation, the false-positive rate was relatively high when $\omega_B = 0$ and 5. It should be noted that the assumptions made for the ω_F and ω_B values in the branch-site test were violated for these cases. The false-positive rate generally increased when ω_F was assumed to be 1. The rate fluctuated according to the ω_B value in a similar manner to the case with $\omega_F = 0.25$, and exceeded 20% when $n = 750$, $\kappa = 4$, and $\omega_B = 5$.

Since positive selection was detected for > 100 genes for most of the parameter sets used in the simulation, a significant fraction of 154 and 233 cases of positive selection detected for the human and chimpanzee branches in the real data analysis (Bakewell et al., 2007) appeared to be false-positives. It should be noted that, in a large-scale analysis, false-positive results may be eliminated by conducting Bonferroni correction and by controlling the FDR. In fact, in the analysis of 13,888 genes from humans, chimpanzees, and macaques, positive selection was detected only for 2 and 21 genes for the human and chimpanzee branches, respectively, by using Bonferroni correction, and for 2 and 59 genes, respectively, by using the FDR (Bakewell et al., 2007). To examine the effectiveness of Bonferroni correction and the FDR in the branch-site test, they were applied to the simulation described above. In both cases, false-positive results were often obtained especially when $\omega_F > 0$ and $\omega_B = 0$ (Table 1), where the assumptions made for the ω_F and ω_B values in the branch-site test were violated. The results did not change to any large extent when different initial ω values were used in the computation for some cases (data not shown).

Interestingly, when the human sequence was compared with the true or inferred ancestral sequence at the interior node for the 99 cases in which positive selection was detected with $n = 450$, $\kappa = 4$, $\omega_F = 0.25$, and $\omega_B = 0.25$, there was a codon site where multiple codon positions were different with $c_N > 1$ in 94 cases (Table 2). For example, for case 8 in Table 2, where positive selection appeared to be erroneously detected for the human branch because $c_N = 2$ (< 4) for the entire sequence, the nonsynonymous substitution was observed only at one codon site. The codons at this site in the human and ancestral sequences were GGA and AAA, respectively, and both of two pathways for the consecutive occurrence of two nucleotide substitutions [AAA (encoding lysine) →

Table 1. Numbers of false-positives obtained by the branch-site test in the computer simulation with 14,000 replications

κ	ω_F	ω_B	n = 450			n = 750		
			Without correction	With Bonferroni correction	With the FDR	Without correction	With Bonferroni correction	With the FDR
4	0.25	0	172	2	2	279	10	17
			(1.23%)[a]	(0.01%)	(0.01%)	(1.99%)	(0.07%)	(0.12%)
	0.25	0.25	99	0	0	138	0	0
			(0.71%)	(0.00%)	(0.00%)	(0.99%)	(0.00%)	(0.00%)
	0.25	1	42	0	0	50	0	0
			(0.30%)	(0.00%)	(0.00%)	(0.36%)	(0.00%)	(0.00%)
	0.25	5	347	0	0	455	0	0
			(2.48%)	(0.00%)	(0.00%)	(3.25%)	(0.00%)	(0.00%)
	1	0	1177	9	22	1192	32	52
			(8.41%)	(0.06%)	(0.16%)	(8.51%)	(0.23%)	(0.37%)
	1	0.25	1141	0	0	1157	0	0
			(8.15%)	(0.00%)	(0.00%)	(8.26%)	(0.00%)	(0.00%)
	1	1	943	0	0	956	0	0
			(6.74%)	(0.00%)	(0.00%)	(6.83%)	(0.00%)	(0.00%)
	1	5	2025	0	0	2949	2	2
			(14.46%)	(0.00%)	(0.00%)	(21.06%)	(0.01%)	(0.01%)
2	0.25	0	208	5	10	381	16	71
			(1.49%)	(0.04%)	(0.07%)	(2.72%)	(0.11%)	(0.51%)
	0.25	0.25	107	0	0	158	0	0
			(0.76%)	(0.00%)	(0.00%)	(1.13%)	(0.00%)	(0.00%)
	0.25	1	48	0	0	58	0	0
			(0.34%)	(0.00%)	(0.00%)	(0.41%)	(0.00%)	(0.00%)
	0.25	5	287	1	1	385	0	0
			(2.05%)	(0.01%)	(0.01%)	(2.75%)	(0.00%)	(0.00%)
	1	0	1400	98	228	1679	441	529
			(10.00%)	(0.70%)	(1.63%)	(11.99%)	(3.15%)	(3.78%)
	1	0.25	1206	0	0	1189	0	0
			(8.61%)	(0.00%)	(0.00%)	(8.49%)	(0.00%)	(0.00%)
	1	1	1031	1	1	1064	1	1
			(7.36%)	(0.01%)	(0.01%)	(7.60%)	(0.01%)	(0.01%)
	1	5	1737	0	0	2528	0	0
			(12.41%)	(0.00%)	(0.00%)	(18.06%)	(0.00%)	(0.00%)

[a]False-positive rate is represented in the parentheses.

GAA (glutamic acid) → GGA (glycine) and AAA → AGA (arginine) → GGA] contained $c_N = 2$. Among the 14,000 replications, the codon site with $c_N > 1$ existed in 125 and 124 replications when the human sequence was compared with the true and inferred ancestral sequences, respectively, indicating that there was a strong correlation between the existence of such a codon site and detection of positive selection ($P = 4.07 \times 10^{-207}$ and 1.01×10^{-207} in Fisher's exact test, respectively). These results suggest that the existence of a codon site with a possibility of occurrence of multiple nonsynonymous substitutions for the foreground branch is a strong indication for the branch-site test to detect positive selection even if positive selection has not actually operated. This association appeared to be weakened as the ω_F and ω_B values became greater (Tables 3 and 4).

Existence of a codon site with $c_N > 1$ for the foreground branch in the positively selected genes detected in the real data analysis In the computer simulation, the branch-site test appeared to detect positive selection when there was a codon site with $c_N > 1$ for

Table 2. c_S and c_N values for the human branch where positive selection was detected by the branch-site test in the computer simulation with $n = 450$, $\kappa = 4$, $\omega_F = 0.25$, and $\omega_B = 0.25$

	Inferred[a]			True[b]				Inferred			True		
Case	c_S	c_N	$c_N > 1$[c]	c_S	c_N	$c_N > 1$	Case	c_S	c_N	$c_N > 1$	c_S	c_N	$c_N > 1$
1	1	1		1	1		51	3	3	✓	3	3	✓
2	2.5	1.5	✓	3.5	1.5	✓	52	0	3	✓	0	3	✓
3	2.5	1.5	✓	2.5	1.5	✓	53	2	3	✓	2	3	✓
4	2.5	1.5	✓	2.5	1.5	✓	54	1	3	✓	2	3	✓
5	5.5	1.5	✓	5.5	1.5	✓	55	1	3	✓	1	3	✓
6	1.2	1.5	✓	1.5	1.5	✓	56	3	3	✓	3	3	✓
7	0.5	1.5	✓	0.5	1.5	✓	57	0	3	✓	0	3	✓
8	3	2	✓	4	2	✓	58	1	3	✓	1	3	✓
9	1	2		1	2		59	2.5	3.5	✓	2.5	3.5	✓
10	1	2	✓	1	2	✓	60	4.5	3.5	✓	4.5	3.5	✓
11	1	2	✓	1	2	✓	61	3.5	3.5	✓	3.5	3.5	✓
12	0	2	✓	0	2	✓	62	2.5	3.5	✓	2.5	3.5	✓
13	1	2	✓	1	2	✓	63	3.5	3.5	✓	3.5	3.5	✓
14	1	2	✓	1	2	✓	64	2.5	3.5	✓	2.5	3.5	✓
15	3	2	✓	3	2	✓	65	1.5	3.5	✓	1.5	3.5	✓
16	6	2	✓	6	2	✓	66	3	3.7	✓	2	4	✓
17	1	2	✓	1	2	✓	67	1.7	3.7		1	4	
18	2	2	✓	2	2	✓	68	1	4	✓	1	4	✓
19	4	2	✓	4	2	✓	69	2	4	✓	2	4	✓
20	0	2	✓	0	2	✓	70	3	4	✓	3	4	✓
21	1.7	2	✓	2	2	✓	71	1	4	✓	1	4	✓
22	1	2	✓	1	2	✓	72	1	4	✓	1	4	✓
23	2.3	2.3	✓	2	2	✓	73	4	4	✓	4	4	✓
24	3.3	2.3	✓	3.5	2.5	✓	74	3	4	✓	3	4	✓
25	4.5	2.5	✓	4.5	2.5	✓	75	4	4	✓	4	4	✓
26	1.5	2.5	✓	1.5	2.5	✓	76	4	4	✓	4	4	✓
27	4.8	2.5	✓	3.5	2.5	✓	77	4.7	4	✓	5	4	✓
28	2.5	2.5	✓	2.5	2.5	✓	78	3	4	✓	3	4	✓
29	0.5	2.5	✓	0.5	2.5	✓	79	3	4	✓	4	4	✓
30	1.5	2.5	✓	1.5	2.5	✓	80	3	4	✓	3	4	✓
31	0.5	2.5	✓	0.5	2.5	✓	81	4	4	✓	4	4	✓
32	3.5	2.5	✓	3.5	2.5	✓	82	1	4	✓	1	4	✓
33	1.5	2.5	✓	1.5	2.5	✓	83	5	4	✓	5	4	✓
34	1.5	2.5	✓	1.5	2.5	✓	84	3	4	✓	3	5	✓
35	2.5	2.5	✓	2.5	2.5	✓	85	0.5	4.5	✓	0.5	4.5	✓
36	1.5	2.5	✓	1.5	2.5	✓	86	2.5	4.5	✓	2.5	4.5	✓
37	3.5	2.5	✓	2.5	2.5	✓	87	3.5	4.5	✓	3.5	4.5	✓
38	3.5	2.5	✓	3.5	2.5	✓	88	3.5	4.5	✓	3.5	4.5	✓
39	2	3	✓	2	2	✓	89	2	5	✓	2	4	✓
40	3	3	✓	2	3	✓	90	2	5	✓	2	5	✓
41	4	3	✓	4	3	✓	91	1	5	✓	1	5	✓
42	5	3	✓	5	3	✓	92	4	5	✓	4	5	✓
43	1	3	✓	1	3	✓	93	1	5	✓	1	5	✓
44	0	3	✓	0	3	✓	94	3	5	✓	3	5	✓
45	1	3	✓	1	3	✓	95	2	5	✓	2	5	✓
46	2	3	✓	4	3	✓	96	1	6	✓	1	6	✓
47	2.7	3	✓	3	3	✓	97	0	6		0	6	
48	1	3	✓	1	3	✓	98	4	6	✓	4	6	✓
49	3	3	✓	3	3	✓	99	0	6		0	6	
50	1	3	✓	1	3	✓							

[a] Inferred ancestral sequence was used.
[b] True ancestral sequence was used.
[c] Checked if a codon site with $c_N > 1$ existed in the sequence.

Table 3. Numbers of replications where c_N was < 4 for the entire sequence and where positive selection was detected for the human branch when the inferred ancestral sequence was used in the computer simulation

				n = 450		n = 750	
κ	ω_F	ω_B	Selection	$c_N \geq 4$	$c_N < 4$	$c_N \geq 4$	$c_N < 4$
4	0.25	0	Detected	55 (46)[a]	117 (80)	115 (93)	164 (78)
			Not detected	684 (4)	13,144 (3)	2751 (6)	10,970 (4)
	0.25	0.25	Detected	32 (30)	67 (64)	102 (88)	36 (34)
			Not detected	684 (3)	13,217 (28)	2895 (19)	10,967 (40)
	0.25	1	Detected	24 (15)	18 (18)	45 (35)	5 (5)
			Not detected	956 (21)	13,002 (58)	3464 (114)	10,486 (76)
	0.25	5	Detected	346 (165)	1 (1)	455 (240)	0 (0)
			Not detected	11,662 (485)	1991 (21)	13,468 (960)	77 (0)
	1	0	Detected	1135 (490)	42 (33)	1189 (728)	3 (2)
			Not detected	10,193 (125)	2630 (4)	12,604 (320)	204 (0)
	1	0.25	Detected	1116 (515)	25 (25)	1155 (742)	2 (2)
			Not detected	10,111 (104)	2748 (2)	12,622 (288)	221 (1)
	1	1	Detected	935 (351)	8 (7)	956 (557)	0 (0)
			Not detected	10,257 (238)	2800 (19)	12,785 (517)	259 (2)
	1	5	Detected	2024 (664)	1 (1)	2949 (1085)	0 (0)
			Not detected	11,782 (774)	193 (3)	11,051 (1285)	0 (0)
2	0.25	0	Detected	48 (31)	160 (88)	178 (106)	203 (76)
			Not detected	849 (2)	12,943 (3)	3096 (12)	10,523 (3)
	0.25	0.25	Detected	36 (31)	71 (68)	110 (104)	48 (45)
			Not detected	794 (3)	13,099 (38)	3030 (17)	10,812 (46)
	0.25	1	Detected	27 (17)	21 (21)	46 (36)	12 (12)
			Not detected	1027 (18)	12,925 (70)	3606 (112)	10,336 (79)
	0.25	5	Detected	283 (146)	4 (0)	385 (214)	0 (0)
			Not detected	11,919 (630)	1794 (19)	13,554 (1101)	61 (0)
	1	0	Detected	1346 (560)	54 (26)	1672 (756)	7 (2)
			Not detected	10,332 (159)	2268 (1)	12,148 (397)	173 (0)
	1	0.25	Detected	1180 (562)	26 (25)	1187 (763)	2 (2)
			Not detected	10,310 (141)	2484 (5)	12,624 (392)	187 (1)
	1	1	Detected	1017 (432)	14 (13)	1063 (609)	1 (1)
			Not detected	10,527 (287)	2442 (20)	12,731 (625)	205 (1)
	1	5	Detected	1734 (623)	3 (0)	2528 (1094)	0 (0)
			Not detected	12,137 (1096)	126 (1)	11,472 (1660)	0 (0)

[a]Number of genes where a codon site with $c_N > 1$ existed in the sequence is represented in the parentheses.

the foreground branch. To examine whether such a codon site existed in the positively selected genes detected in the real data analysis, 13,888 genes from humans, chimpanzees, and macaques (Bakewell et al., 2007) were re-analyzed, and the results are summarized in Table 5. Among the 154 and 233 genes for which positive selection was detected for the human and chimpanzee branches, such a codon site existed in 127 and 197 genes, respectively, suggesting that there was a strong correlation between the existence of such a codon site and detection of positive selection ($P = 2.94 \times 10^{-223}$ and $< 1.63 \times 10^{-322}$ in Fisher's exact test, respectively). In addition, among the 154 and 233 genes, $c_N < 4$ in 53 and 82 genes for the human and chimpanzee branches, respectively. These results suggest that a significant fraction of positively selected genes that have been detected for the human and chimpanzee branches are false-positives. Among the 13,888 genes, the number of genes containing a codon site with $c_N > 1$ for the chimpanzee branch (284) was greater than that for the human branch (210) ($P = 0.00087$ in χ^2 test). Although the reason for this observation remains to be elucidated, this may be a cause for

Table 4. Numbers of replications where c_N was < 4 for the entire sequence and where positive selection was detected for the human branch when the true ancestral sequence was used in the computer simulation

				n = 450		n = 750	
κ	ω_F	ω_B	Selection	$c_N \geq 4$	$c_N < 4$	$c_N \geq 4$	$c_N < 4$
4	0.25	0	Detected	56 (47)[a]	116 (79)	115 (93)	164 (78)
			Not detected	690 (4)	13,138 (2)	2777 (6)	10,944 (3)
	0.25	0.25	Detected	34 (31)	65 (63)	103 (89)	35 (33)
			Not detected	701 (3)	13,200 (27)	2951 (22)	10,911 (39)
	0.25	1	Detected	22 (13)	20 (19)	42 (32)	8 (7)
			Not detected	778 (19)	13,180 (50)	2902 (92)	11,048 (76)
	0.25	5	Detected	80 (19)	267 (31)	201 (46)	254 (24)
			Not detected	714 (22)	12,939 (50)	2752 (62)	10,793 (61)
	1	0	Detected	1136 (488)	41 (31)	1190 (727)	2 (2)
			Not detected	10,221 (125)	2602 (2)	12,610 (320)	198 (0)
	1	0.25	Detected	1120 (517)	21 (21)	1155 (739)	2 (2)
			Not detected	10,244 (106)	2615 (2)	12,645 (283)	198 (0)
	1	1	Detected	938 (343)	5 (4)	956 (544)	0 (0)
			Not detected	10,437 (246)	2620 (18)	12,817 (524)	227 (3)
	1	5	Detected	1916 (320)	109 (10)	2938 (539)	11 (1)
			Not detected	9467 (275)	2508 (16)	10,868 (473)	183 (2)
2	0.25	0	Detected	49 (32)	159 (87)	180 (107)	201 (75)
			Not detected	860 (2)	12,932 (3)	3119 (11)	10,500 (3)
	0.25	0.25	Detected	36 (31)	71 (68)	109 (102)	49 (46)
			Not detected	843 (2)	13,050 (36)	3119 (20)	10,723 (46)
	0.25	1	Detected	26 (14)	22 (19)	46 (35)	12 (11)
			Not detected	875 (17)	13,077 (61)	3126 (101)	10,816 (77)
	0.25	5	Detected	56 (17)	231 (15)	186 (35)	199 (17)
			Not detected	761 (24)	12,952 (55)	3122 (88)	10,493 (52)
	1	0	Detected	1347 (559)	53 (26)	1672 (757)	7 (2)
			Not detected	10,362 (157)	2238 (1)	12,153 (390)	168 (0)
	1	0.25	Detected	1180 (557)	26 (25)	1187 (760)	2 (2)
			Not detected	10,481 (138)	2313 (5)	12,638 (399)	173 (1)
	1	1	Detected	1018 (411)	13 (11)	1063 (585)	1 (1)
			Not detected	10,722 (281)	2247 (13)	12,754 (601)	182 (2)
	1	5	Detected	1656 (312)	81 (9)	2522 (565)	6 (0)
			Not detected	10,082 (409)	2181 (19)	11,302 (643)	170 0)

[a]Number of genes where a codon site with $c_N > 1$ existed in the sequence is represented in the parentheses.

Table 5. Numbers of genes where c_N was < 4 for the entire sequence and where positive selection was detected for the human and chimpanzee branches in the real data analysis of genes from humans, chimpanzees, and macaques

	Human branch			Chimpanzee branch		
Selection	$c_N \geq 4$	$c_N < 4$	Total	$c_N \geq 4$	$c_N < 4$	Total
Detected	101 (79)[a]	53 (48)	154 (127)	151 (126)	82 (71)	233 (197)
Not detected	1561 (53)	12,173 (30)	13,734 (83)	1405 (45)	12,250 (42)	13,655 (87)
Total	1662 (132)	12,226 (78)	13,888 (210)	1556 (171)	12,332 (113)	13,888 (284)

[a]Number of genes where a codon site with $c_N > 1$ existed in the sequence is represented in the parentheses.

a greater number of positively selected genes detected for the chimpanzee branch than for the human by the branch-site test (Arbiza et al., 2006; Bakewell et al., 2007).

DISCUSSION

In the computer simulation, the false-positive rate was found to inflate when the assumptions made for ω_F and ω_B in the branch-site test were violated, where background branches were strongly negatively selected (ω_B = 0) or positively selected (ω_B = 5). In addition, under a similar condition, false-positive results were often obtained even when Bonferroni correction was conducted and the FDR was controlled in a large-scale analysis. False-positive results were also obtained even when c_N for the foreground branch was smaller than the minimum value required for detecting positive selection. It should be noted that, in the LRT of the branch-site test, $2\Delta\ln L$ was assumed to follow χ_1^2 despite the fact that $2\Delta\ln L$ was asymptotically distributed as an equal mixture of a point mass on 0 and χ_1^2 to make the test conservative. The false-positive rate would be even greater if the correct distribution was used.

The existence of a codon site with a possibility of occurrence of multiple nonsynonymous substitutions for the foreground branch appeared to be mostly sufficient for explaining false-positives for the branch-site test, especially when the ω_F and ω_B values were relatively small. In fact, such a codon site existed in most of the positively selected genes detected in the computer simulation and in the real data analysis. Since the branch-site test is intended to detect a group of codon sites with $\omega_2 > 1$, genes containing such codon sites are likely to be detected as positively selected. However, such a codon site can be generated by chance without positive selection, and the branch-site test tends to detect positive selection even in this case, as shown in the computer simulation. In addition, in the real data, the frequency of codon sites with the occurrence of substitutions at multiple codon positions appears to be high (Dayhoff et al., 1972; Averof et al., 2000; Whelan and Goldman, 2004). These observations indicate that many of the positively selected genes detected by the branch-site test that contain a codon site with $c_N > 1$ for the foreground branch are false-positives in the real data analysis.

The present study was motivated by discussions with Masatoshi Nei. The author thanks Masatoshi Nei and two anonymous reviewers for valuable suggestions and comments. The author is indebted to Margaret A. Bakewell and Jianzhi Zhang for providing the data set of 13,888 orthologous trios of protein-coding genes from humans, chimpanzees, and macaques that was analyzed in Bakewell et al. (2007). The present study was supported by the National Institutes of Health grant GM020293 to Masatoshi Nei and KAKENHI 17770007 to Y. S.

REFERENCES

Arbiza, L., Dopazo, J., and Dopazo, H. (2006) Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. PLoS Comput. Biol. **2**, e38.

Averof, M., Rokas, A., Wolfe, K. H., and Sharp, P. M. (2000) Evidence for a high frequency of simultaneous double-nucleotide substitutions. Science **287**, 1283–1286.

Bakewell, M. A., Shi, P., and Zhang, J. (2007) More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc. Natl. Acad. Sci. USA **104**, 7489–7494.

Dayhoff, M. O., Eck, R. V., and Park, C. M. (1972) A model of evolutionary change in proteins. In: Atlas of Protein Sequence and Structure, volume 5. (ed.: M. O. Dayhoff), pp. 89–99, National Biomedical Research Foundation, Washington, D. C.

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

Hughes, A. L., and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**, 167–170.

Jiang, C., and Zhao, Z. (2006) Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms. Genomics **88**, 527–534.

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

Rhesus Macaque Genome Sequencing and Analysis Consortium (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science **316**, 222–234.

Rosenberg, M. S., Subramanian, S., and Kumar, S. (2003) Patterns of transitional mutation biases within and among mammalian genomes. Mol. Biol. Evol. **20**, 988–993.

Storey, J. D. (2002) A direct approach to false discovery rates. J. R. Stat. Soc. B **64**, 479–498.

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

Whelan, S., and Goldman, N. (2004) Estimating the frequency of events that cause multiple-nucleotide changes. Genetics **167**, 2027–2043.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24**, 1586–1591.

Yang, Z., and Nielsen, R. (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. **19**, 908–917.

Zhang, J., Nielsen, R., and 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.