

False-Positive Selection Identified by ML-Based Methods: Examples from the *Sig1* Gene of the Diatom *Thalassiosira weissflogii* and the *tax* Gene of a Human T-cell Lymphotropic Virus

Yoshiyuki Suzuki* and Masatoshi Nei†

*Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Japan; and †Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University

Sexually induced gene 1 (*Sig1*) in the centric diatom *Thalassiosira weissflogii* is considered to encode a gamete recognition protein. Sorhannus (2003) analyzed nucleotide sequences of *Sig1* using parsimony analysis and the maximum-likelihood (ML)-based Bayesian method for inferring positive selection at single amino acid sites and reported that positively selected sites were detected by the latter method but not by the former. He then concluded that for this type of study, the ML-based method is more reliable than parsimony analysis. Here we show that his results apparently represent false-positive cases of the ML-based method and that there is no solid evidence that this gene contains positively selected sites. We further demonstrate that in the *tax* gene of human T-cell lymphotropic virus type I (HTLV-I), all codon sites, including invariable sites, can be inferred as positively selected sites by the ML-based method. These observations indicate that the ML-based method may produce many false-positive sites. One of the main reasons for the occurrence of false positives is that in the ML-based method, codon sites are grouped into several categories, with different nonsynonymous/synonymous rate ratios (ω), on a purely statistical basis, and positive selection is inferred indirectly by examining whether the average ω for each category is greater than 1. In parsimony analysis, however, the evolutionary change of nucleotides at each codon site is examined. For this reason, parsimony-based methods rarely produce false positives and are safer than ML-based methods for detecting positive selection at individual codon sites, although a large number of sequences are necessary.

Introduction

Diatoms are photosynthetic protists, which are classified into centric and pennate diatoms, according to whether they have a radial or bilateral symmetry. They undergo sexual and asexual reproduction. Sexually induced gene 1 (*Sig1*) was identified as one of the genes in centric diatoms (*Thalassiosira* spp.) whose transcription level was highly up-regulated during sexual reproduction (Armbrust 1999). The SIG1 protein is localized to the extracellular matrix and considered to be involved in gamete recognition. There appear to be multiple loci of *Sig1* in the genome with multiple alleles at each locus, some of which are pseudogenes (Armbrust and Galindo 2001). The sequence divergence of *Sig1* between related species is reasonably high, but positive selection was not detected when the rate of nonsynonymous nucleotide substitution per nonsynonymous site (r_N) was compared with that of synonymous substitution per synonymous site (r_S) for the entire coding region, r_N/r_S ($= \omega$) being 0.12 (Armbrust and Galindo 2001).

It is, however, possible that positive selection operates only for a subset of codon sites. Sorhannus (2003) analyzed *Sig1* sequences of *Thalassiosira weissflogii* using the parsimony-based method (Suzuki and Gojobori 1999) and the ML-based Bayesian method (Yang et al. 2000) for detecting positive selection at individual codon sites and concluded that the latter method detected positively selected sites, whereas the former method did

not. However, the ML-based method is known to produce many false-positive results for positive selection, whereas the parsimony-based method rarely does so (Suzuki and Nei 2002). It is, therefore, possible that the ML results obtained by Sorhannus (2003) are, in fact, false positives.

Sorhannus (2003) commented that “the results obtained by likelihood analysis of HLA data by Suzuki and Nei (2001) appear to be problematical as simpler models had much higher likelihood values than more general models and multiple runs led to many different sets of parameter estimates.” After publication of our paper, N. Goldman, R. Nielsen, and Z. Yang (personal communication) informed us that the version 3.0a of the computer program PAML, which we used, contained an inaccurate computing algorithm. However, even when we used the new version, PAML 3.12, the disconcerting results mentioned by Sorhannus (2003) did not completely disappear (Suzuki and Nei, unpublished data).

The purpose of this paper is to show that the results obtained by Sorhannus (2003) are apparently false positives caused by using an unreliable phylogenetic tree and that there is no compelling evidence of positive selection in *Sig1*. In addition, false-positive selection observed in the *tax* gene in human T-cell lymphotropic virus type I (HTLV-I) will be presented as an extreme example. The reasons that ML-based methods produce many false positives are discussed.

Materials and Methods

The same sets of nucleotide sequences of *Sig1* from *Thalassiosira weissflogii* as those used by Sorhannus (2003) were obtained from the author. Sorhannus (2003) analyzed two data sets: a “large data set” and a “small

Key words: Positive selection, parsimony, likelihood, *Thalassiosira weissflogii*, sexually induced gene 1, human T-cell lymphotropic virus type I, *tax*.

E-mail: yossuzuk@lab.nig.ac.jp.

Mol. Biol. Evol. 21(5):914–921. 2004

DOI:10.1093/molbev/msh098

Advance Access publication March 10, 2004

Table 1
The c_N and c_S Values at Negatively Selected Codon Sites in *Sigl* Inferred by the Parsimony-Based Method in Sorhannus (2003)

Position	Large Data Set			Small Data Set		
	p -Distance	d_S -Distance	Sorhannus	p -Distance	d_S -Distance	Sorhannus
8	0/2 ^a	0/1	0/2	0/1	0/1	0/2
10	1/1	1/1	1/1	1/0	1/0	1/0
34	0/2	0/1	0/2	0/1	0/1	0/2
47	0/1	0/1	0/1	0/0	0/0	0/0
49	0/1	0/1	0/1	0/0	0/0	0/0
76	0/1	0/1	0/1	0/0	0/0	0/0
86	0/1	0/1	0/1	0/0	0/0	0/0
116	0/1	0/1	0/1	0/0	0/0	0/0
181	0/1	0/1	0/1	0/0	0/0	0/0

data set.” The large data set consisted of 34 sequences whose accession numbers are AF374490, AF374492 to AF374498, AF374500 to AF374503, AF374505 to AF374510, AF374511, AF374513 to AF374520, AF374523, AF374524, and AF374526 to AF374530 (Armbrust and Galindo 2001). The small data set was a subset of the large data set and contained 25 sequences whose accession numbers are underlined. Note that these sequences were derived from multiple loci of the *Sigl* gene. Each sequence consisted of 186 codon sites.

A multiple sequence alignment was made for each data set using ClustalW version 1.81 (Thompson, Higgins, and Gibson 1994). In each data set, sequences were quite homogeneous, and there were no alignment gaps. Positive selection was inferred at each codon site by the parsimony-based method with ADAPTSITE version 1.3 (Suzuki, Gojobori, and Nei 2001) and the ML-based method as implemented in PAML version 3.13 (Yang 1997). The detailed procedures of the statistical methods are explained by Suzuki and Gojobori (1999) and Yang et al. (2000). In both methods, the phylogenetic tree of sequences used is assumed to be known and has to be preassigned. Sorhannus (2003) made a composite tree of the bootstrap consensus trees of *Sigl* and β -*tubulin*, which were originally constructed by Armbrust and Galindo (2001). However, the tree was highly multifurcative and appeared to be unreliable (see Supplementary Material online). We, therefore, constructed the new trees using the neighbor-joining (NJ) method (Saitou and Nei 1987). To examine the effect of topological difference on the inference of positive selection, we constructed two trees for each data set using two different evolutionary distances; that is, p -distance (proportion of different nucleotide sites) and d_S -distance (estimated number of synonymous nucleotide substitutions by Nei and Gojobori’s [1986] method) (see Supplementary Material online). It is known that the ML-based method sometimes produces different estimates of ω at a given codon site, depending on the input ω value, because of multiple local maxima on the likelihood surface. For this reason, we used 0.4, 3.14, and 4 as the input ω values. (Only the results with the highest log-likelihood [lnL] values were presented in tables 2 and 3.) The significance (confidence) level (cutoff point) for inferring positive selection was 0.05 (0.95) for both parsimony-based and ML-based methods.

Results

Reanalysis of *Sigl* Gene Sequences

Using parsimony analysis, Sorhannus (2003) found positively selected sites in neither the large nor the small data sets. However, he identified nine sites (positions 8, 10, 34, 47, 49, 76, 86, 116, and 181) as negatively selected, although the data set used for this analysis was not clearly mentioned. Our reanalysis of both data sets by the parsimony-based method did not detect any positively selected sites, whether p -distance or d_S -distance was used for constructing trees. However, we did not find negatively selected sites, either. To examine the reason why we failed to infer the nine negatively selected sites, we estimated the numbers of synonymous (c_S) and nonsynonymous (c_N) nucleotide substitutions for the entire tree at each codon site by the maximum-parsimony (MP) method. Because the sequences analyzed were closely related, the estimates of c_S and c_N obtained must be quite reliable (Saitou 1989). As shown in table 1, the c_S and c_N values at these sites were at most 1 or 2 when the p -distance and d_S -distance trees were used. These numbers are obviously too small to detect any type of selection. Curiously, the c_S and c_N values did not change significantly when we used the Sorhannus tree.

By contrast, the ML analysis of Sorhannus (2003) inferred positive selection for four sites (positions 4, 42, 52, and 149) for the large data set and seven sites (positions 4, 9, 42, 52, 119, 149, and 182) for the small data set. In reanalysis of the same data sets, we first used the p -distance tree. The results obtained by using different input ω values (0.4, 3.14, and 4) were essentially the same for all models in both large and small data sets (the difference in the lnL values being smaller than 1). For the large data set, both models M3 and M8 indicated existence of a group of positively selected sites (table 2). However, none of these sites had a Bayesian posterior probability greater than 95%. In addition, the likelihood ratio test (LRT) did not show that M3 and M8 fit the data better than the null models M0 and M7, respectively. In the small data set, a positively selected group was again identified, but the LRT did not show that models M3 and M8 were better than M0 and M7, respectively. In short, we failed to identify positively selected sites in both large and small data sets.

Table 2
Positively Selected Sites in *Sigl* Inferred by the ML-Based Method for the p -Distance Tree^a

Data Set	Model	lnL	Proportion ^b	Selection ^c	Positively Selected Sites
Large	M0	-1259.973		$\omega = 0.381$	None ^d
	M3	-1257.991	$p_0 = 0.956$	$\omega_0 = 0.286$	None
			$p_1 = 0.038$	$\omega_1 = 3.020$	
			$p_2 = 0.007$	$\omega_2 = 3.020$	
M7	-1258.410		$p = 0.227$ $q = 0.356$	Not allowed ^e	
M8	-1257.992	$p_0 = 0.956$ $p_1 = 0.044$	$p = 39.864$ $q = 99.000$	None	
			$\omega_1 = 3.040$		
Small	M0	-1050.323		$\omega = 0.949$	None
	M3	-1047.408	$p_0 = 0.682$	$\omega_0 = 0.000$	4, 9, 10, 37, 42, 43, 52, 60, 63, 83, 84, 95, 98, 100, 119, 126, 137, 139, 144, 149, 159, 178, 182
			$p_1 = 0.252$	$\omega_1 = 3.159$	
			$p_2 = 0.066$	$\omega_2 = 3.159$	
	M7	-1049.437		$p = 0.008$ $q = 0.003$	Not allowed
	M8	-1047.408	$p_0 = 0.682$ $p_1 = 0.318$	$p = 0.003$ $q = 6.029$	4, 9, 10, 37, 42, 43, 52, 60, 63, 83, 84, 95, 98, 100, 119, 126, 137, 139, 144, 149, 159, 178, 182
$\omega_1 = 3.159$					

^a M3 and M8 were not judged to fit the data better than M0 and M7 by the LRT, respectively.
^b p_0 , p_1 , and p_2 indicate the proportions of groups 0, 1, and 2 in each model, respectively.
^c ω_0 , ω_1 , and ω_2 indicate the ω values of groups 0, 1, and 2 in each model, respectively. p and q are beta parameters.
^d No codon site belonged to the category with $\omega > 1$ with the posterior probability of >95%.
^e Positively selected sites are not allowed to exist in M7.

This situation changed dramatically when we used the d_S -distance tree. In each data set, existence of a group of positively selected sites was identified, and this group included many codon sites (table 3). In addition, both M3 and M8 were judged to fit the data better than M0 and M7,

respectively. For the large data set, the initial ω values of 0.4, 3.14, and 4 all indicated that five sites (positions 4, 37, 42, 52, and 149) and four sites (positions 37, 42, 52, and 149) were positively selected with M3 and M8, respectively. For the small data set, 23 sites (positions 4, 9,

Table 3
Positively Selected Sites in *Sigl* Inferred by the ML-Based Method for the d_S -Distance Tree^a

Data Set	Model	lnL	Proportion ^b	Selection ^c	Positively Selected Sites
Large	M0	-1365.485		$\omega = 0.542$	None ^d
	M3	-1337.456	$p_0 = 0.914$	$\omega_0 = 0.215$	4, 37, 42, 52, 149
			$p_1 = 0.053$	$\omega_1 = 3.014$	
			$p_2 = 0.033$	$\omega_2 = 9.370$	
M7	-1350.185		$p = 0.013$ $q = 0.026$	Not allowed ^e	
M8	-1337.469	$p_0 = 0.955$ $p_1 = 0.045$	$p = 0.354$ $q = 0.834$	37, 42, 52, 149	
			$\omega_1 = 8.229$		
Small	M0	-1142.887		$\omega = 1.576$	None
	M3	-1110.543	$p_0 = 0.745$	$\omega_0 = 0.000$	4, 9, 10, 37, 42, 43, 52, 60, 63, 83, 84, 95, 98, 100, 119, 126, 137, 139, 144, 149, 159, 178, 182
			$p_1 = 0.223$	$\omega_1 = 3.803$	
			$p_2 = 0.032$	$\omega_2 = 33.788$	
	M7	-1136.258		$p = 0.003$ $q = 0.005$	Not allowed
	M8	-1111.000	$p_0 = 0.941$ $p_1 = 0.059$	$p = 0.022$ $q = 0.019$	4, 9, 10, 37, 42, 43, 52, 60, 63, 83, 84, 95, 98, 100, 119, 126, 137, 139, 144, 149, 159, 178, 182^f
$\omega_1 = 23.398$					

^a M3 and M8 were judged to fit the data better than M0 and M7 by the LRT, respectively (bold faced).
^b p_0 , p_1 , and p_2 indicate the proportions of groups 0, 1, and 2 in each model, respectively.
^c ω_0 , ω_1 , and ω_2 indicate the ω values of groups 0, 1, and 2 in each model, respectively. p and q are beta parameters.
^d No codon site belonged to the category with $\omega > 1$ with the posterior probability of >95%.
^e Positively selected sites are not allowed to exist in M7.
^f When the initial values of $\omega = 0.4$ and 3.14 were used, the same 23 sites as those for M3 were identified with $\ln L = -1115.173$. However, when the initial value of $\omega = 4$ was used, only the four sites underlined were identified as positively selected, and the $\ln L$ was -1111.000 .

Table 4
Quantities Qualifying the Accuracy of Phylogenetic Trees Used for *SigI*

Quantity	Large Data Set			Small Data Set		
	<i>p</i> -Distance	<i>d_S</i> -Distance	Sorhannus	<i>p</i> -Distance	<i>d_S</i> -Distance	Sorhannus
TL ^a	77	91	87	40	53	49
S _A ^b	0.445	0.563	0.523	0.227	0.312	0.300
CI ^c	0.948	0.802	0.839	0.900	0.679	0.735
RI ^d	0.983	0.922	0.939	0.789	0.105	0.316
RC ^e	0.932	0.739	0.788	0.711	0.071	0.232

^a Total tree length (total number of nucleotide substitutions for the entire tree).

^b Total number of nucleotide substitutions per codon for the entire tree.

^c Consistency index.

^d Retention index.

^e Rescaled consistency index.

10, 37, 42, 43, 52, 60, 63, 83, 84, 95, 98, 100, 119, 126, 137, 139, 144, 149, 159, 178, and 182) were identified as positively selected for both M3 and M8 when the initial ω values of 0.4 and 3.14 were used. However, when $\omega = 4$ was used as the initial value, the results for M8 changed drastically, and the selected sites were now 37, 42, 52, and 149 only. (The same four sites were obtained when we tried the initial ω values of 5, 6, and 7.) Therefore, the detection of positively selected sites is dependent on the initial ω value, as was indicated by Suzuki and Nei (2001), and the results obtained can be very different, depending on the initial ω value. In the present case, initial $\omega = 4$ gave a higher $\ln L$ value (-1111.000) than that for initial $\omega = 0.4$ or 3.14 (-1115.173).

At any rate, we have obtained three different results from the same data sets: the *p*-distance tree, the *d_S*-distance tree, and the Sorhannus tree. These different results must be caused by differences in the trees used, because there is no other difference. Previously, Yang et al. (2000) stated that inference of positively selected sites does not seem to be sensitive to the assumed topology. In the present data sets, however, this is not the case. To examine which tree and which computational results are most reliable, we computed the $\ln L$ values. The $\ln L$ value for the large data set with M3 was -1257.99 for the *p*-distance tree, -1337.46 for the *d_S*-distance tree, and -1308.00 for the Sorhannus tree. Similarly, the $\ln L$ value was highest for *p*-distance tree and lowest for *d_S*-distance tree for all models in both large and small data sets. We also computed the total tree length (TL), consistency index (CI), retention index (RI), and rescaled consistency index (RC) for the three trees. All these indices indicated that the *p*-distance tree was best and the *d_S*-distance tree was worst (table 4). In addition, the *p*-distance tree was judged to fit the data better than other two trees by the tests of Templeton (1983) and Kishino and Hasegawa (1989) ($P < 0.05$), whereas the latter two trees were not significantly different from each other. It is also known that *p*-distance is generally most reliable for constructing NJ trees of closely related sequences (Takahashi and Nei 2000). To examine whether there was a more reliable tree than the *p*-distance tree, we constructed phylogenetic trees using the MP method with PAUP* version 4.0b10 (Swofford 1998) for both the large and small data sets. In addition, we examined the best-fit model of nucleotide substitution among all possible models currently available with MODELTEST version

3.06 (Posada and Crandall 1998) and constructed trees using the NJ and ML methods assuming that model. It was found that the best-fit model was the Kimura (1980) model for both large and small data sets, and the topologies of the MP, NJ, and ML trees obtained were all identical to that of the *p*-distance tree. These results strongly suggest that the *p*-distance tree is most reliable for both large and small data sets. Therefore, the result for the *p*-distance tree, in which no positively selected site was inferred, appears to be most reliable, and positive selection identified by Sorhannus (2003) appears to be false positives caused by the unreliable tree used.

Why did the ML-based method produce many false positives in *SigI*?

To answer this question, we computed the c_S and c_N values by parsimony methods for all 186 codon sites. The results obtained for the sites where positive selection was inferred are presented in table 5. For other sites, c_N and c_S were mostly 0 or 1. As expected, the "selected sites" generally have high c_N values. If a codon site has a c_N value of 3 or higher, the site is "selected" except in position 4 in the *p*-distance tree. In the small data set, however, even the sites with $c_N/c_S = 2/0$ or $2/1$ can be "selected" in the Sorhannus tree. In the *d_S*-distance tree, even the sites with $c_N/c_S = 1/0$ or $1/1$ can be "selected," but there are also sites with $c_N/c_S = 5/0$ and $6/0$.

To have some idea about the expected ratio of c_S and c_N under no selection, we computed the average number of potential synonymous sites per sequence (S) and that of potential nonsynonymous site per sequence (N) (Nei and Kumar 2000) for the large data set. We obtained $S = 128$ and $N = 430$. Therefore, c_N and c_S are expected to occur with a ratio of 3.4 and 1 under no selection. If we conduct the test of selective neutrality at individual codon site separately by the parsimony-based method (using the standard binomial probability), the minimum value of c_N required at a given codon site to be inferred as positively selected (with $c_S = 0$) at the 5% significance level (one-tailed test) is 12. This indicates that even the site with $c_N/c_S = 6/0$ in table 5 can be explained by chance alone. For this reason, a large number of sequences are required in this method (Suzuki and Gojobori 1999).

In the ML-based method, codon sites are grouped into two or more categories with different ω values, and the

Table 5
The c_N and c_S Values at Positively Selected Codon Sites Inferred by the ML-Based Method in *Sig1*

Position	Large Data Set			Small Data Set		
	p -Distance	d_S -Distance	Sorhannus	p -Distance	d_S -Distance	Sorhannus
4	3/0 ^a	3/0	3/0	2/0	2/0	2/0
9	2/1	2/1	2/1	2/0	2/0	2/0
10	1/1	1/1	1/1	1/0	1/0	1/0
37	1/1	5/1	1/1	1/0	5/0	1/0
42	2/0	6/0	4/0	2/0	6/0	4/0
43	1/0	1/0	1/0	1/0	1/0	1/0
52	2/0	3/0	3/0	2/0	3/0	3/0
60	1/1	1/1	1/1	1/0	1/0	1/0
63	1/0	2/0	2/0	1/0	2/0	2/0
83	2/0	2/0	2/0	2/0	2/0	2/0
84	1/0	1/0	1/0	1/0	1/0	1/0
95	1/0	1/0	1/0	1/0	1/0	1/0
98	1/0	1/0	1/0	1/0	1/0	1/0
100	1/1	1/1	1/1	1/0	1/0	1/0
119	1/0	2/0	2/0	1/0	2/0	2/0
126	1/0	1/0	1/0	1/0	1/0	1/0
137	1/0	1/0	1/0	1/0	1/0	1/0
139	1/1	1/1	1/1	1/1	1/1	1/1
144	1/0	1/0	1/0	1/0	1/0	1/0
149	1/0	4/0	4/0	1/0	4/0	4/0
159	1/0	1/0	1/0	1/0	1/0	1/0
178	1/0	1/0	1/0	1/0	1/0	1/0
182	1/1	2/1	2/1	1/1	2/1	2/1

^a The c_N and c_S values are given before and after the slash sign, respectively. The results are bold faced when positive selection is inferred.

null hypothesis of $\omega = 1$ is tested indirectly for the group with $\omega > 1$. Because the sites with high ω values are grouped into the $\omega > 1$ category, this method is more efficient than the parsimony method in detecting selection if the high ω values are caused by selection. In practice, however, ω is affected by stochastic errors, and it is possible that most high ω values are caused by random errors. For example, if $c_S = 0$ by chance at a given codon site with $c_N > 0$, ω becomes theoretically infinite (table 5). Similarly, ω can easily be inflated if c_N becomes large or c_S becomes small by chance. If this happens, the ω value for the $\omega > 1$ group may again become significantly higher than 1, but this does not mean that the codon sites involved have been subjected to positive selection. This is the main reason why the ML-based method can produce many false positives.

In table 5, we have seen that the d_S -distance and Sorhannus trees generated many codon sites with high c_N values and produced many positively selected sites in comparison with the p -distance tree. Because the actual process of identification of positively selected sites is quite complicated, it is difficult to see the exact relationships between c_N values and the number of positively selected sites. However, the reason the number of positively selected sites is larger in the d_S -distance and Sorhannus trees than in the p -distance tree seems to be that when there are many codon sites with large c_N and small c_S values, the average ω value for the group of so-called "selected sites" can be reasonably high even if more sites with lower c_N values are included. If this argument is right, one would expect that a poor tree that generates many sites with high c_N values would give more "selected sites" than a better tree. This is indeed what we observed in tables 2, 3, and 5.

For example, the c_N/c_S value at position 4 for the large data set was 3/0, regardless of the trees assumed (table 5). However, this site was inferred as positively selected for the d_S -distance and Sorhannus trees but not for the p -distance tree, probably because the c_N values at another sites (e.g., positions 37, 42, and 149) were inflated for the former trees.

However, if positive selection at individual sites can be falsely identified because of stochastic errors, it would happen irrespective of the topology used. In the following, we show a striking example in which positive selection was inferred even at invariable codon sites under the assumption of a reliable tree.

Striking Example of Inferred Selected Sites: the *tax* Gene of HTLV-I

Twenty nucleotide sequences of the *tax* gene of HTLV-I were extracted from the international nucleotide sequence database. The accession numbers of these sequences were AB045401, AB45410, AB045425, AB045442, AB045481, AB045482, AB045486, AB045490, AB045514, AB045519, AB045520, AB045528, AB045541, AB045546 to AB045549, AB045558, AB045559, and AB45639 (Furukawa et al. 2001). Each sequence consisted of 181 codon sites that were not overlapped with the open reading frame of the *rex* gene. These sequences were highly homogeneous, and there were no alignment gaps. The total branch length per codon site for the entire tree (S_A ; Anisimova, Bielawski, and Yang's S) was 0.128. Parsimony-based and ML-based methods were used for inferring positively selected sites. The phylogenetic tree was constructed by the NJ method with p -distance. The

Table 6
Results from the ML Analysis of the *tax* Gene (181 Codons)^a

Model	lnL	Proportion ^b	Selection ^c	Positively Selected Sites
M0	-892.030		$\omega = 4.870$	All
M3	-892.030	$p_0 = 0.000$ $p_1 = 0.000$ $p_2 = 1.000$	$\omega_0 = 0.000$ $\omega_1 = 2.454$ $\omega_2 = 4.870$	All
M7	-895.509		$p = 0.524$ $q = 0.001$	Not allowed ^d
M8	-892.030	$p_0 = 0.000$ $p_1 = 1.000$	$p = 0.543$ $q = 2.047$ $\omega_1 = 4.870$	All

^a M8 was judged to fit the data better than M7 by the LRT (bold faced). Although M3 did not fit the data better than M0, M0 also indicated that all codon sites were positively selected (bold faced).

^b p_0 , p_1 , and p_2 indicate the proportions of groups 0, 1, and 2 in each model, respectively.

^c ω_0 , ω_1 , and ω_2 indicate the ω values of groups 0, 1, and 2 in each model, respectively. p and q are beta parameters.

^d Positively selected sites are not allowed to exist in M7.

tree obtained was a star phylogeny, which was obviously the most reliable tree for these sequences because all mutations were singletons.

Parsimony analysis did not detect any positively selected sites, because c_S and c_N values were all small and either 0 or 1. Surprisingly, however, the ML-based method indicated that all codon sites were positively selected (table 6). That is, M3 and M8 both indicated that a group of positively selected codon sites existed with probability 1, and all codon sites were inferred to belong to this group with a Bayesian posterior probability of 1. In addition, M8 was judged to fit the data better than M7 by the LRT. Interestingly, the lnL value for M3 was very close to that for M0, but it was shown that all codon sites were positively selected in both models. Surprisingly, 158 out of a total of 181 codon sites analyzed were invariable among all the aligned sequences.

This unexpected result was obtained apparently because the c_N/c_S value was 1/0 for 21 variable sites, 0/1 for the two remaining variable sites, and 0/0 for all the invariable sites. This results in that the weighted average of $(c_N/N)/(c_S/S)$, that is, $\bar{\omega} = (\sum_i c_{Ni}/N)/(\sum_i c_{Si}/S) = 3.4$ for the entire sequence, where i refers to the i th codon site. With this $\bar{\omega}$ value, one may infer that all sites are positively selected. Of course, this is a simplified explanation of this unusual observation, and the actual procedure for detecting selected sites in the ML-based model is much more complicated.

Incidentally, the possibility of occurrence of $\omega > 1$ for all codon sites when closely related sequences are analyzed was previously indicated by Anisimova, Bielawski, and Yang (2002) in a computer simulation. From this simulation, they suggested that the ML-based method should be used only when the number of sequences used (T) is greater than 6 and the total number of nucleotide substitutions per codon for the entire tree (S_A) is greater than 0.11. In the present case $T = 20$ and $S_A = 0.128$, so this condition is satisfied. Yet, we observed a case of $\omega > 1$ for all sites. However, the real problem is that false positives can occur even when both T and S_A are large

(Suzuki and Nei 2002). Anisimova, Bielawski, and Yang's simulation was not intended to study false positives, but the example data set of $T = 17$ and $S_A = 0.38$ in their figure 1B indicates that false positives occurred with this data set, because the accuracy of predicting positively selected sites was lower than the cutoff point P when $P > 0.85$.

Some Anomalous Observations

Finally, it should be pointed out that essentially the same lnL values and same parameter estimates were obtained for different models M3 and M8 in tables 2 and 6. Similar results were also obtained for models M3 and M8 in table 1 of Sorhannus (2003). Furthermore, even M0 gave the same results as those of M3 and M8 in table 6. These unexpected results were apparently obtained because the different distributions of ω assumed among different sites converged to the same one in different models. In addition, in PAML, the beta distribution of ω in M8 is approximated by 10 discrete categories of ω , and these categories apparently converged to the three categories of M3 in table 2 and to one category of M0 in table 6.

Discussion

The ML-based method appears to be attractive to many experimentalists because this method easily generates positively selected sites. In fact, a number of authors (e.g., Yang et al. 2000, Swanson and Vacquier 2002; Bailey et al. 2003; Miller 2003) reported findings of positively selected codon sites in several different proteins. However, this method is known to produce false-positive results even when no positively selected sites actually exist (Suzuki and Nei 2002) or when positively selected sites and negatively selected sites are mixed (Anisimova, Bielawski, and Yang 2002). It is therefore important to examine any inferred selective sites carefully by using various statistical methods. Note that the parsimony method proposed by Suzuki and Gojobori (1999) employs the standard statistical approach for testing selection at individual sites. The power of this method is certainly low, unless a large number of sequences are used. However, this is what one would expect in the test of selection at individual codon sites.

By contrast, the ML-based method is intended to identify a group of codon sites for which positive selection might operate simultaneously. If there is indeed such a group as in the case of major histocompatibility complex (MHC) genes, use of this method is justified. However, if high ω values are generated by random errors, this method is expected to give false positives. Furthermore, the general applicability of Goldman and Yang's (1994) codon substitution model, which is the basis of the ML-based model, has been questioned (Nei and Kumar 2000). As far as we know, there are no empirical data to support this model. Because it is very difficult to distinguish between true positives and false positives only from the sequence analysis, it is important to exercise caution in the interpretation of the results obtained by this method. Obviously, the final proof of positive selection rests on experimental verification, as was done with some genes

(e.g., Jermann et al. 1995; Zhang, Zhang, and Rosenberg 2002; Shi and Yokoyama 2003). As long as the conclusion is drawn only from the sequence analysis without conducting experiments, a more conservative parsimony-based method is safer for detecting positive selection at single amino acid sites.

Detection of a significant excess of nonsynonymous substitutions is often taken as evidence that the protein under consideration undergoes adaptive evolution. In host-defense genes or immunogenic genes, excess nonsynonymous substitutions caused by positive selection seem to occur continuously to avoid parasitic attack or immune surveillance, as in the cases of MHC genes (Hughes and Nei 1988, 1989) and immunogenic genes in human immunodeficiency and influenza viruses (Bush et al. 1999; Suzuki and Gojobori 1999). For these genes, the statistical methods considered here would be useful if a large number of sequences are used. In many genes, however, both advantageous and deleterious mutations caused by nonsynonymous substitutions may occur at different times or nearly at the same time, and the overall function of a gene may remain more or less the same in long-term evolution. In this case, positive or negative selection at individual codon sites may not have significant effects on the evolution of a gene. For this reason, it is important to distinguish between excess nonsynonymous nucleotide substitution and positive selection for enhancing gene function. In other words, detection of excess nonsynonymous substitutions at some codon sites does not necessarily mean the detection of positive selection for gene function. By contrast, only one or two amino acid changes at a few codon sites may affect the function of a gene drastically, as in the case of alligator hemoglobin (Perutz 1983) or vertebrate color vision genes (Yokoyama and Radlwimmer 2001). In this case, it would be difficult to detect positive selection by the statistical methods considered in this paper.

Acknowledgments

The authors thank Ulf Sorhannus for providing us with sequence and tree files used in his study. We are also grateful to two anonymous reviewers for their valuable comments. This study was supported by NIH grant GM 29332.

Literature Cited

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.

Armbrust, E. V. 1999. Identification of a new gene family expressed during the onset of sexual reproduction in the centric diatom *Thalassiosira weissflogii*. *Appl. Environ. Microbiol.* **65**:3121–3128.

Armbrust, E. V., and H. M. Galindo. 2001. Rapid evolution of a sexual reproduction gene in centric diatoms of the genus *Thalassiosira*. *Appl. Environ. Microbiol.* **67**:3501–3513.

Bailey, X., R. Leroy, S. Canrey, O. Collin, F. Zal, A. Toulmond, and D. Jollivet. 2003. The loss of the hemoglobin HsS-binding function in annelids from sulfide-free habitats reveals

molecular adaptation driven by Darwinian positive selection. *Proc. Natl. Acad. Sci. USA* **100**:5885–5890.

Bush, R. M., C. A. Bender, K. Subbarao, N. J. Cox, and W. M. Fitch. 1999. Predicting the evolution of human influenza A. *Science* **286**:1921–1925.

Furukawa, Y., R. Kubota, M. Tara, S. Izumo, and M. Osame. 2001. Existence of escape mutant in HTLV-1 *tax* during the development of adult T-cell leukemia. *Blood* **97**:987–993.

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

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

———. 1989. Nucleotide substitution at major histocompatibility complex II loci: evidence for overdominant selection. *Proc. Natl. Acad. Sci. USA* **86**:958–962.

Jermann, T. M., J. G. Opitz, J. Stackhouse, and S. A. Benner. 1995. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* **374**:57–59.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**:111–120.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topology from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* **29**:170–179.

Miller, S. R. 2003. Evidence for the adaptive evolution of the carbon fixation gene *rbcL* during diversification in temperature tolerance of a clade of hot spring cyanobacteria. *Mol. Ecol.* **12**:1237–1246.

Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**:418–426.

Nei, M., and S. Kumar. 2000. *Molecular evolution and phylogenetics*. Oxford University Press, New York.

Perutz, M. F. 1983. Species adaptation in a protein molecule. *Mol. Biol. Evol.* **1**:1–28.

Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**:817–818.

Saitou, N. 1989. A theoretical study of the underestimation of branch lengths by the maximum parsimony principle. *Sys. Zool.* **38**:1–6.

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

Shi, Y., and S. Yokoyama. 2003. Molecular analysis of the evolutionary significance of ultraviolet vision in vertebrates. *Proc. Natl. Acad. Sci. USA* **100**:8308–8313.

Sorhannus, U. 2003. The effect of positive selection on a sexual reproduction gene in *Thalassiosira weissflogii* (Bacillariophyta): results obtained from maximum likelihood and parsimony-based methods. *Mol. Biol. Evol.* **20**:1326–1328.

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

Suzuki, Y., T. Gojobori, and M. Nei. 2001. ADAPTSITE: detecting natural selection at single amino acid sites. *Bioinformatics* **17**:660–661.

Suzuki, Y., and M. Nei. 2001. Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* **18**:2179–2185.

———. 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.

Swanson, W. J., and V. D. Vacquier. 2002. The rapid evolution of reproductive proteins. *Nat. Rev. Genet.* **3**:137–144.

- Swofford, D. L. 1998. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Mass.
- Takahashi, K., and M. Nei. 2000. Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. *Mol. Biol. Evol.* **17**:1251–1258.
- Templeton, A. R. 1983. Phylogenetic inference from restriction cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution* **37**:221–244.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673–4680.
- Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**:555–556.
- Yang, Z., R. Nielsen, N. Goldman, and A. M. K. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**:431–449.
- Yokoyama, S., and F. B. Radlwimmer. 2001. The molecular genetics and evolution of red and green color vision in vertebrates. *Genetics* **158**:1697–1710.
- Zhang, J., Y. P. Zhang, and H. F. Rosenberg. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat. Genet.* **30**:411–415.

Naruya Saitou, Associate Editor

Accepted January 6, 2004