

## Small-Sample Tests of Episodic Adaptive Evolution: A Case Study of Primate Lysozymes

Jianzhi Zhang, Sudhir Kumar, and Masatoshi Nei

Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University

Positive Darwinian selection at the molecular level is often studied by comparing the number of synonymous nucleotide substitutions per synonymous site ( $d_S$ ) and the number of nonsynonymous substitutions per nonsynonymous site ( $d_N$ ) between homologous gene sequences, and a  $t$ -test with an infinite number of degrees of freedom is usually used for determining the statistical significance of the difference between  $d_S$  and  $d_N$  (Hughes and Nei 1988; Kumar, Tamura, and Nei 1993). An assumption required for this test is that the sample size (number of substitutions between the sequences) is so large that  $d_S$  and  $d_N$  are approximately normally distributed. However, it is unclear how well the  $t$ -test performs when the sample size is small and whether there is a better way of testing positive selection. In this letter, we address these issues by using a recently published data set of primate lysozyme sequences (Messier and Stewart 1997) as an example.

In higher vertebrates, lysozyme is usually expressed in macrophages, tears, saliva, mammalian milk, and avian egg white as a host defense protein to fight against invading bacteria. In ruminants, colobine monkeys, and hoatzins (an avian species), however, lysozyme has been recruited in stomachs for digestion of bacteria passing through the guts to extract the nutrients assimilated by the bacteria. Recently, Messier and Stewart (1997) presented an interesting way of analyzing of adaptive evolution by comparing  $d_S$  and  $d_N$  for the inferred ancestral sequences of primate lysozyme genes. They concluded that there was an episode of positive Darwinian selection in each of the ancestral branches of colobines (branch *a* of fig. 1A) and hominoids (branch *b* of fig. 1A) and that these episodes were followed by negative selection (episodic evolution). While the occurrence of positive selection in branch *a* is justifiable to explain the evolution of foregut fermentation in colobines, the occurrence of positive selection in branch *b* is biologically puzzling.

Messier and Stewart (1997) first inferred the ancestral nucleotide sequences of the lysozyme genes for the interior nodes of the primate tree (fig. 1A). They then used the inferred sequences to compute  $d_S$  and  $d_N$  for each branch and tested the difference between  $d_S$  and  $d_N$ . They obtained  $d_S = 0.00722$ ,  $SE(d_S) = 0.00726$ ,  $d_N = 0.03374$ , and  $SE(d_N) = 0.01150$  for branch *a* of figure 1 by using Li's (1993) method,

Key words: positive selection, episodic evolution, lysozyme, primates, sample size.

Address for correspondence and reprints: Jianzhi Zhang, 322 Mueller Laboratory, Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, Pennsylvania 16802. E-mail: jxz128@psu.edu.

*Mol. Biol. Evol.* 14(12):1335–1338, 1997

© 1997 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

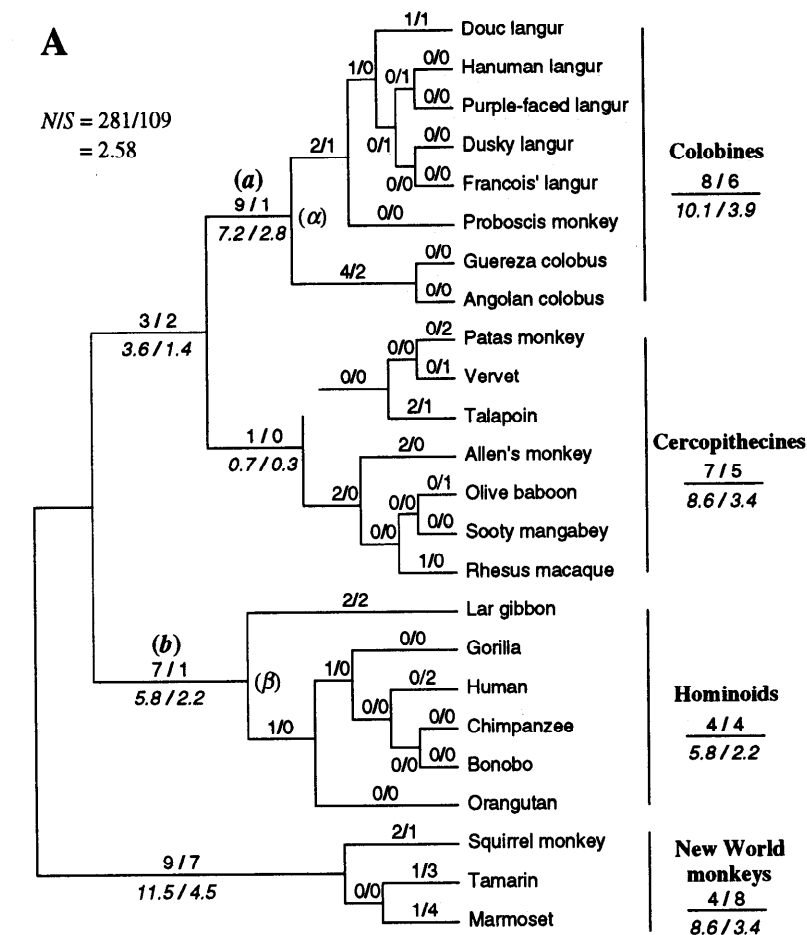
where  $SE(d_S)$  and  $SE(d_N)$  are the standard errors of  $d_S$  and  $d_N$ , respectively. In this case, the test statistic ( $t_s$ ) becomes

$$t_s = \frac{d_N - d_S}{\sqrt{[SE(d_N)]^2 + [SE(d_S)]^2}} = 1.95.$$

Under the assumption that the  $t_s$  follows the  $t$  distribution with an infinite number of degrees of freedom or the standard normal distribution, Messier and Stewart (1997) found that  $d_N$  was significantly greater than  $d_S$  ( $t_s > 1.65$  and  $P < 0.05$  in the one-tailed test) and concluded that positive selection operated for the branch. For branch *b*, a similar  $t_s$  value (1.82) was obtained, and the same conclusion was reached.

In the present case, however, the numbers of synonymous ( $s$ ) and nonsynonymous ( $n$ ) substitutions are so small (fig. 1A) that the statistic  $t_s$  is unlikely to follow the  $t$  distribution, and the  $t$ -test may reject the null hypothesis of neutral evolution more often than expected by chance. We therefore conducted a computer simulation to examine the actual distribution of  $t_s$  when the expected numbers of synonymous and nonsynonymous substitutions per site are equal (i.e., neutral evolution). In this simulation, we used the inferred lysozyme gene sequence (390 nt long) at the ancestral node of branch *a* and introduced random substitutions to generate a descendant sequence with an expected number of substitutions per site equal to 0.02. This value was chosen to represent the observed lengths of branches *a* and *b* approximately. We used an expected transition/transversion ratio ( $R$ ) of 2, which was close to the observed value. Once a descendant sequence was obtained, the sequence was compared with the ancestral sequence, and  $d_S$  and  $d_N$  were computed using Li's (1993) method. We then computed the test statistic  $t_s$ . This was repeated 10,000 times, and the empirical distribution of  $t_s$  was obtained (fig. 2). Comparison of this distribution with the  $t$  (or normal) distribution indicates that the former distribution is skewed and that a  $t_s$  value corresponding to a 5% significance level of the  $t$  distribution actually has a type I error of 11%. Therefore, the use of  $t_s$  will reject the null hypothesis of neutral evolution two times as often as required. The distribution of  $t_s$  approaches that of  $t$  as the sequence length and the level of sequence divergence increase, as expected (data not shown). The large-sample  $t$ -test may be used when both the numbers of synonymous and nonsynonymous substitutions exceed about 10.

Since the large-sample test is not applicable to the primate lysozyme sequences, we need a new way of testing neutral evolution. In this data set, all the sequences are so closely related that  $s$  and  $n$  can simply be counted for each branch and compared with their expected numbers under the hypothesis of neutral evolu-



**B Test of positive selection**

	Non.	Syn.	Prob.
<b>Branch a</b>			
Changes	9 ( $n$ )	1 ( $s$ )	
No changes	272 ( $N-n$ )	108 ( $S-s$ )	<b>0.18</b>
<b>Branch b</b>			
Changes	7 ( $n$ )	1 ( $s$ )	
No changes	274 ( $N-n$ )	108 ( $S-s$ )	<b>0.30</b>

**C Test of episodic evolution**

	Numbers of substitutions		
	Non.	Syn.	Prob.
<b>Colobines</b>			
Branch a	9	1	
Desc. lineages	8	6	<b>0.10</b>
<b>Hominoids</b>			
Branch a	7	1	
Desc. lineages	4	4	<b>0.14</b>

FIG. 1.—Tests of episodic adaptive evolution of primate lysozymes. **A**, Numbers of synonymous ( $s$ ) and nonsynonymous ( $n$ ) nucleotide substitutions per sequence for each branch of the phylogenetic tree of primate lysozyme genes. The ancestral nucleotide sequences at the interior nodes were inferred by the Bayesian method (Yang, Kumar, and Nei 1995; Zhang and Nei 1997). The values of  $n$  and  $s$  are given as  $n/s$  for each branch (above the line), whereas their expected numbers [ $(n + s)N/(N + S)$  and  $(n + s)S/(N + S)$ , respectively] are given in italics below the line. The  $n$  and  $s$  values for a group of primate species represent the sums of nonsynonymous and synonymous substitutions, respectively, for all the branches involved. For branch  $b$ ,  $n/s$  was either 7/1 or 7.5/0.5, depending on the weight for alternative pathways (Nei and Gojobori 1986), but this did not affect our conclusion. There was no ambiguity in determining the  $n$  and  $s$  values for other branches. Use of the ancestral sequences inferred by the parsimony method (Fitch 1971) did not change our conclusion either. **B**, Tests of positive selection. **C**, Tests of episodic evolution. Abbreviations: non, nonsynonymous; syn, synonymous; prob, tail probabilities in Fisher's exact test of homogeneity; desc, descendant.

tion (fig. 1A). Using Ina's (1995) method with  $R = 2$ , we also computed the numbers of synonymous ( $S$ ) and nonsynonymous ( $N$ ) sites for each sequence. The  $S$  and  $N$  values obtained were approximately 109 and 281, respectively, for all the sequences. Under the null hypothesis of neutral evolution, i.e., equal rates of synonymous

and nonsynonymous substitution, the ratio of  $n/s$  is expected to be equal to  $N/S$ . Using Fisher's exact test, we found that neutral evolution cannot be rejected for either branch  $a$  ( $P = 0.18$ ) or branch  $b$  ( $P = 0.30$ ) (fig. 1B). We have used various values of  $R$  (from 0.5 to 5) in the computation but reached the same conclusion.

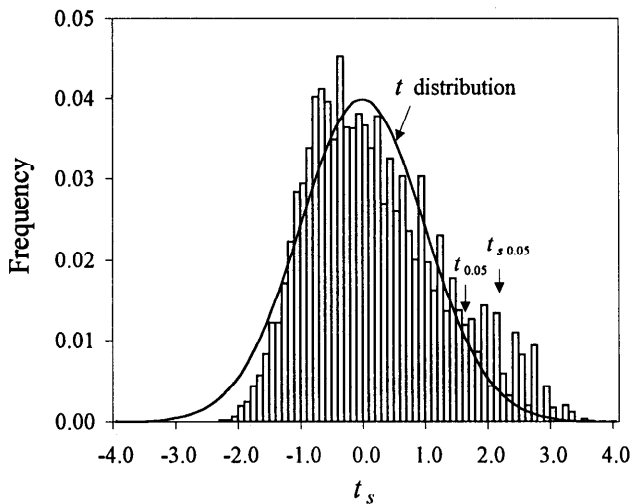


FIG. 2.—The distribution of the statistic  $t_s$  (histograms) and the theoretical distribution of  $t$  (curve). The critical values at the 5% significance level (one-tailed test) are indicated by  $t_{0.05}$  and  $t_{s, 0.05}$  for the distributions of  $t$  and  $t_s$ , respectively.

Fisher's exact test can also be used to examine whether  $d_N$  is significantly greater than  $d_S$  for the intergroup comparisons of colobine, cercopithecine, and hominoid sequences. For example, we have  $n = 19$  and  $s = 4$  between the most recent common ancestors of the colobines and the hominoids ( $\alpha$  and  $\beta$  in fig. 1A). The probability of this event is 0.18 under the hypothesis of neutral evolution, so neutrality cannot be rejected. Similar results were obtained for the other two intergroup comparisons. It is worth mentioning that, apart from the use of the large-sample test, the differences between our results and those reported by Messier and Stewart (1997) are also caused by underestimation of the standard errors of  $d_S$  and  $d_N$  in their analysis. This underestimation happened because they used Kimura's (1980) model to estimate  $d_S$  and  $d_N$  (computer program by Li 1993) but Jukes and Cantor's (1969) model (as implemented in the *SEND* program [Nei and Jin 1989]) to estimate the standard errors. This also made their tests too liberal in rejecting the neutral evolution hypothesis.

The hypothesis of positive selection followed by negative selection (episodic evolution) can also be tested by comparing  $n$  and  $s$  for an ancestral branch and its descendants. In the case of colobines,  $n/s$  is 9/1 for the ancestral branch  $a$  and 8/6 for the descendant lineages (i.e., all the branches linking the colobine species; see fig. 1A). Fisher's exact test shows that these two ratios are not significantly different ( $P = 0.10$ ; fig. 1C). A similar conclusion was obtained for the hominoids ( $P = 0.14$ ). Therefore, Messier and Stewart's (1997) data are not sufficient to establish episodic evolution statistically. It should be noted that in both Messier and Stewart's test and ours,  $n$  and  $s$  were treated as if they were observed. This treatment is justifiable in the present case because the accuracy (posterior probability) of the inferred ancestral sequences was over 99.5% on average.

We have seen that proper statistical tests do not support Messier and Stewart's (1997) conclusion about episodic adaptive evolution of primate lysozymes. How-

ever, this does not mean that there have been no amino acid substitutions driven by positive selection. If we consider the fact that most amino acid residues of a protein are subject to purifying selection, the relatively high ratio of  $d_N/d_S$  in this gene may be an indication of positive selection, although the ratio is not significantly higher than 1. Nevertheless, it is not clear how the function of lysozyme has been changed by the amino acid substitutions in relation to the foregut fermentation. Since the primary function of lysozyme is to fight against invading bacteria, the enzyme may show a higher rate of amino acid substitution than average proteins even without involvement in foregut fermentation (Murphy 1993). It appears that to identify amino acid substitutions involved in the evolution of foregut fermentation, it is necessary to produce ancestral proteins by site-directed mutagenesis and to examine the functional change of lysozyme in the evolutionary process, as was done in the case of artiodactyl ribonuclease (Jermann et al. 1995).

### Acknowledgments

We thank Caro-Beth Stewart for clarifying some details of her analysis and Xun Gu and Tatyana Sitnikova for thoughtful discussions. This work was supported by NIH and NSF research grants to M.N.

### LITERATURE CITED

- FITCH, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.* **20**:406–416.
- 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.
- INA, Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J. Mol. Evol.* **40**:190–226.
- 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.
- JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–123 in H. N. MUNRO, ed. *Mammalian protein metabolism*. Academic Press, New York.
- 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.
- KUMAR, S., K. TAMURA, and M. NEI. 1993. MEGA: molecular evolutionary genetics analysis. Version 1.02. The Pennsylvania State University, University Park.
- LI, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitutions. *J. Mol. Evol.* **36**:96–99.
- MESSIER, W., and C.-B. STEWART. 1997. Episodic adaptive evolution of primate lysozymes. *Nature* **385**:151–154.
- MURPHY, P. M. 1993. Molecular mimicry and the generation of host defense protein diversity. *Cell* **72**:823–826.
- 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 L. JIN. 1989. Variances of the average numbers of nucleotide substitutions within and between populations. *Mol. Biol. Evol.* **6**:290–300.

YANG, Z., S. KUMAR, and M. NEI. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**:1641–1650.

ZHANG, J., and M. NEI. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J. Mol. Evol.* **44**:S139–S146.

DAN GRAUR, reviewing editor

Accepted September 5, 1997