

Letter to the Editor

Nucleotide Substitution Rate Estimation in Enterobacteria: Approximate and Maximum-Likelihood Methods Lead to Similar Conclusions

Nick G. C. Smith and Adam Eyre-Walker

Centre for the Study of Evolution and School of Biological Sciences, University of Sussex, Brighton, England

Synonymous mutations are “silent” with regard to the amino acid sequence of a protein, but a wealth of evidence indicates that, at least in species with large effective population sizes, synonymous mutations are subject to translational selection (Akashi and Eyre-Walker 1998). One line of evidence for translational selection has been the perceived negative correlation between codon bias and synonymous substitution rates (K_S), reported in both *Drosophila* (Shields et al. 1988) and enterobacteria (Sharp and Li 1987a). The reasoning behind this interpretation of the correlation between K_S and codon bias is straightforward. Translational selection means that certain codons are favored over others, so selection causes an increased codon bias with a preponderance of favored codons, which means that most synonymous mutations are selectively disfavored and so selection also reduces K_S .

The selective interpretation of the correlation between K_S and codon bias in *Drosophila* has been questioned recently on both theoretical and methodological grounds. It seems that translational selection is weak relative to genetic drift in *Drosophila* (Akashi and Schaeffer 1997), but theoretical models show that when selection for codon usage is weak there is unlikely to be a strong relationship between codon bias and K_S (McVean and Charlesworth 1999). Furthermore, the reported relationship between codon bias and K_S in *Drosophila* (Shields et al. 1988) has recently been shown to be a methodological artifact of approximate substitution rate estimation methods which fail to account fully for biased codon usage and transition/transversion biases (Dunn, Bielawski, and Yang 2001). When more realistic maximum-likelihood (ML) methods are used to estimate substitution rates, no relationship between K_S and codon bias is found (Dunn, Bielawski, and Yang 2001).

Given these new findings for *Drosophila*, we investigated the methodological sensitivity of the correlation between K_S and codon bias in enterobacteria. We used a set of 128 aligned protein-coding sequences from *Escherichia coli* and *Salmonella typhimurium* (Eyre-Walker and Bulmer 1995). Codon bias was measured by the codon adaptation index (CAI), which gives a measure of how closely codon usage approaches that found in highly expressed genes. CAI was calculated only for the *E. coli* genes using the method of Sharp and Li

(1987b) with modifications suggested by Bulmer (1988). The estimation of substitution rates by ML analysis was performed with the PAML package (Yang 1997), which accounts for the transition/transversion rate bias and codon usage bias (Goldman and Yang 1994; Yang and Nielsen 1998). Codon usage was modeled by empirical estimation of the 61 codon frequencies, which was shown by likelihood ratio tests to provide a significant improvement over a model using base frequencies at the three codon positions for all genes (results not shown). Substitution rates were also computed using two approximate methods: that of Nei and Gojobori (1986), hereinafter termed the NG method, and that of Li, Wu, and Luo (1985), hereinafter termed the LWL method. The NG method was used to enable direct comparison with previous studies of methodological sensitivity (Bielawski, Dunn, and Yang 2000; Dunn, Bielawski, and Yang 2001), and the LWL method was used to provide direct comparison with the first report of a correlation between K_S and codon bias in enterobacteria (Sharp and Li 1987a). We removed from consideration those genes which appeared to be approaching saturation at synonymous sites (ML $K_S > 3$) and those genes for which the approximate methods failed to provide substitution rate estimates, which left 99 genes.

Estimates of K_S were found to be sensitive to methodology, with ML rates being about 50% higher than both NG and LWL rates (mean ML $K_S = 1.36$, mean NG $K_S = 0.88$, mean LWL $K_S = 0.85$; Mann-Whitney U -test, $P < 0.001$ for both ML versus approximate comparisons). These differences are due to the failures of the NG and LWL methods to account fully for transition/transversion biases and biased codon usage. We can separate these two effects by considering alternative simplified ML models. In the first case, the transition/transversion ratio was fixed at 1, which led to an 84% increase in mean ML K_S , from 1.36 to 2.50. In the second case, the codon frequencies were assumed to be equal (1/61), which led to a 47% decrease in mean ML K_S , from 1.36 to 0.72. Thus, the two failings of the NG and LWL methods cancel each other out to some extent, but the effect of biased codon usage predominates, as previously found for mammalian genes (Bielawski, Dunn, and Yang 2000).

However, the negative correlation between K_S and CAI was insensitive to methodology (see fig. 1), although the approximate methods gave stronger correlations (Spearman's rank correlation; ML $r^2 = 0.15$ and $P < 0.001$; NG $r^2 = 0.34$ and $P < 0.001$; LWL $r^2 = 0.35$ and $P < 0.001$). Hence, given that the negative correlation between codon bias and K_S in *E. coli* is not a methodological artifact, how can we explain the relationship?

Key words: translational selection, codon bias, methodology, maximum likelihood, synonymous substitution rates.

Address for correspondence and reprints: Nick Smith, Centre for the Study of Evolution and School of Biological Sciences, University of Sussex, Brighton BN1 9QG, United Kingdom. E-mail: n.g.c.smith@sussex.ac.uk.

Mol. Biol. Evol. 18(11):2124–2126. 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

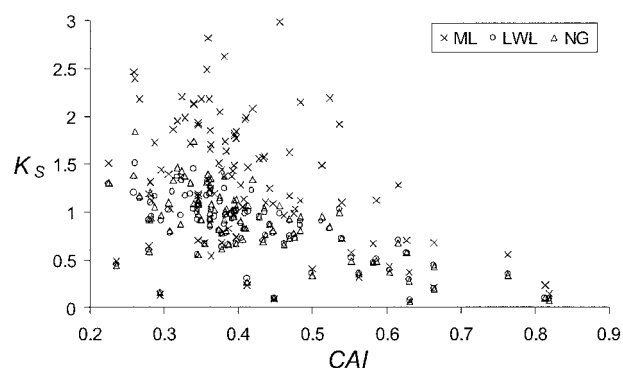


FIG. 1.—The relationship between synonymous substitution rates (K_S) and codon bias (CAI) for 99 enterobacterial genes, with K_S estimated using a maximum-likelihood (ML) method and the approximate Nei and Gojobori (1986) (NG) and Li, Wu, and Luo (1985) (LWL) methods.

Translational selection appears to be weak relative to genetic drift in *E. coli* (Hartl, Moriyama, and Sawyer 1994; Smith and Eyre-Walker 2001), and so there is little reason to expect that weak selection is the cause of such a strong relationship between K_S and CAI (McVean and Charlesworth 1999). Indeed, it is very unlikely that weak selection can explain such a wide range of K_S values, from 0.065 to >3 (see fig. 1). A more likely explanation for the negative correlation between K_S and codon bias is variation in mutation rates mediated by expression levels, as first shown by Berg and Martelius (1995). Genes which are expressed at high levels have low mutation rates (due to transcription-mediated repair; see Selby and Sancar 1993), which are the primary determinant of K_S , and these genes are also under greater translational selection and thus also have high codon bias. This mutational explanation is also supported by the work of Eyre-Walker and Bulmer (1995), who showed that K_S is lower in highly expressed genes even for those amino acids which show little change in codon bias across expression levels.

Although we find no methodological sensitivity in the relationship between K_S and codon bias in *E. coli*, that does not mean that substitution rate methodology is of no concern in the study of enterobacterial evolution. As described above, K_S is about 50% higher with the ML method than with the NG and LWL methods. We also considered two other molecular-evolution statistics which have been shown to be sensitive to methodology: the correlation between K_S and the nonsynonymous substitution rate K_A , and the ratio K_A/K_S (Bielawski, Dunn, and Yang 2000; Dunn, Bielawski, and Yang 2001). A significant positive correlation between K_A and K_S was produced by all three methods (Spearman's rank correlation; ML $r^2 = 0.22$ and $P < 0.001$; NG $r^2 = 0.37$ and $P < 0.001$; LWL $r^2 = 0.39$ and $P < 0.001$), but estimates of K_A/K_S were sensitive to methodology, with the ML ratio being about 50% lower than the NG and LWL ratios (mean ML $K_A/K_S = 0.033$, mean NG $K_A/K_S = 0.047$, mean LWL $K_A/K_S = 0.047$; Mann-Whitney U -test, $P < 0.001$ for both ML versus approximate comparisons). The differences in K_A/K_S between ML and approximate methods are almost wholly due to the dif-

ferences in K_S , which is more sensitive than K_A to changes in methodology (see Li 1993).

The methodological sensitivity of K_S and K_A/K_S estimates is potentially important. Accurate estimation of K_S is required for estimates of mutation rates on an evolutionary timescale, and the K_A/K_S ratio has been used both as a test of positive selection (Yang and Bielawski 2000) and as a measure of constraint at nonsynonymous sites (Keightley and Eyre-Walker 2000). Indeed, the methodological sensitivity of K_S and K_A/K_S estimates in enterobacteria corroborates previous reports of the importance of substitution rate estimation methods in *Drosophila* and mammals (Bielawski, Dunn, and Yang 2000; Dunn, Bielawski, and Yang 2001). The approach taken in those previous studies, and also adopted here, is that differences between the results of ML and approximate methods are probably due to biases in the approximate methods. This viewpoint is supported by the results obtained using the alternative simplified ML models. However, we note two important respects in which it is incorrect to presume that ML results are "right" and approximate results are "wrong."

First, ML methods are unbiased only if the correct model is used. We performed some model testing to show that empirical estimation of codon frequencies was justified, but even the most complex models can only ever approximate biological reality. One failing of the ML model used here is its failure to account fully for variation among sites within genes. The only variation between sites allowed in the ML model used here is between synonymous and nonsynonymous sites. However, there appears to be considerable heterogeneity in mutation rates within enterobacterial genes, with important consequences for the estimation of mutation rates (Berg 1999), and there are probably also heterogeneous selection pressures at amino acid sites which affect K_A/K_S estimates (Yang et al. 2000). However, such variation among sites is unlikely to affect the primary finding of this paper, the confirmation of the negative correlation between K_S and codon bias, unless the levels of among-sites variation differ in a particular way among genes (unless variation among sites is greatest for those genes for which we have estimated the K_S to be lowest).

Second, ML methods often invoke more complex models than approximate methods, which means that ML methods tend to require more data, or, to put it another way, approximate methods will tend to have the advantage of lower variance relative to ML methods for the same data (note that the K_S -CAI correlation is stronger with the NG and LWL methods than with the ML method). This advantage of approximate methods may justify the cost of their higher bias in some cases (Takahashi and Nei 2000). In view of such differences between ML and approximate methods, the results presented here strengthen the findings for *Drosophila* (Dunn, Bielawski, and Yang 2001). An alternative explanation for the methodological sensitivity of the correlation between K_S and codon bias in *Drosophila* is that the greater variance of the ML estimates of K_S masks the true underlying relationship, which is correctly revealed by the approximate methods. However, our dem-

onstration of a significant negative correlation between K_S and CAI in enterobacteria using an ML method indicates that the increased variance of ML methods does not always mask evolutionary patterns.

Acknowledgments

We thank Ziheng Yang for advice on the use of PAML. N.G.C.S. was funded by the BBSRC, and A.E.-W. was funded by the Royal Society.

LITERATURE CITED

- AKASHI, H., and A. EYRE-WALKER. 1998. Translational selection and molecular evolution. *Curr. Opin. Genet. Dev.* **8**: 688–693.
- AKASHI, H., and S. W. SCHAEFFER. 1997. Natural selection and the frequency distributions of “silent” DNA polymorphism in *Drosophila*. *Genetics* **146**:295–307.
- BERG, O. G. 1999. Synonymous nucleotide divergence and saturation: effects of site-specific variations in codon bias and mutation patterns. *J. Mol. Evol.* **48**:398–407.
- BERG, O. G., and M. MARTELIUS. 1995. Synonymous substitution-rate constants in *Escherichia coli* and *Salmonella typhimurium* and their relationship to gene expression and selection pressure. *J. Mol. Evol.* **41**:449–456.
- BIELAWSKI, J. P., K. A. DUNN, and Z. H. YANG. 2000. Rates of nucleotide substitution and mammalian nuclear gene evolution: approximate and maximum-likelihood methods lead to different conclusions. *Genetics* **156**:1299–1308.
- BULMER, M. 1988. Are codon usage patterns in unicellular organisms determined by mutation selection balance? *J. Evol. Biol.* **1**:15–26.
- DUNN, K. A., J. P. BIELAWSKI, and Z. YANG. 2001. Substitution rates in *Drosophila* nuclear genes: implications for translational selection. *Genetics* **157**:295–305.
- EYRE-WALKER, A., and M. BULMER. 1995. Synonymous substitution rates in enterobacteria. *Genetics* **140**:1407–1412.
- GOLDMAN, N., and Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**:725–736.
- HARTL, D. L., E. N. MORIYAMA, and S. A. SAWYER. 1994. Selection intensity for codon bias. *Genetics* **138**:227–234.
- KEIGHTLEY, P. D., and A. EYRE-WALKER. 2000. Deleterious mutations and the evolution of sex. *Science* **290**:331–333.
- LI, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36**: 96–99.
- LI, W. H., C. I. WU, and C. C. LUO. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2**:150–174.
- MCVEAN, G. A. T., and B. CHARLESWORTH. 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. *Genet. Res.* **74**:145.
- NEI, M., and T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**:418–426.
- SELBY, C. P., and A. SANCAR. 1993. Molecular mechanism of transcription-repair coupling. *Science* **260**:53–58.
- SHARP, P. M., and W. H. LI. 1987a. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol. Biol. Evol.* **4**:222–230.
- . 1987b. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**:1281–1295.
- SHIELDS, D. C., P. M. SHARP, D. G. HIGGINS, and F. WRIGHT. 1988. Silent sites in *Drosophila* genes are not neutral—evidence of selection among synonymous codons. *Mol. Biol. Evol.* **5**:704–716.
- SMITH, N. G. C., and A. EYRE-WALKER. 2001. Why are translationally sub-optimal synonymous codons used in *Escherichia coli*? *J. Mol. Evol.* (in press).
- 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.
- 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.
- YANG, Z. H., and J. P. BIELAWSKI. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15**: 496–503.
- YANG, Z. H., and R. NIELSEN. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* **46**:409–418.

HOWARD OCHMAN, reviewing editor

Accepted July 12, 2001