Adjusting for Selection on Synonymous Sites in Estimates of Evolutionary Distance

Aaron E. Hirsh, Hunter B. Fraser,† and Dennis P. Wall*,[1]*

*Department of Biological Sciences, Stanford University, Stanford, California; †Department of Molecular and Cell Biology, University of California, Berkeley

Evolution at silent sites is often used to estimate the pace of selectively neutral processes or to infer differences in divergence times of genes. However, silent sites are subject to selection in favor of preferred codons, and the strength of such selection varies dramatically across genes. Here, we use the relationship between codon bias and synonymous divergence observed in four species of the genus *Saccharomyces* to provide a simple correction for selection on silent sites.

Introduction

The number of synonymous nucleotide substitutions per synonymous site, or *dS*, is central to a variety of analyses in the study of molecular evolution, including the construction of gene genealogies (Nei 1996; Langkjaer et al. 2003), the investigation of mutational processes (Birdsell 2002), and the estimation of the nature and intensity of selection (Suzuki and Gojobori 1999; Yang 2001). In all of these applications, a common assumption is that *dS* is a measure of evolutionary divergence caused by the selectively neutral processes of mutation and drift. For instance, in the estimation of the strength of selection, the ratio of nonsynonymous to synonymous rates of substitution, *dN*/*dS*, is viewed as a measure of the departure from neutrality caused by selection on nonsynonymous sites. However, work on prokaryotes, fungi, *C. elegans*, *Drosophila*, and *Arabidopsis* has shown that selection also operates on synonymous sites, favoring codons that allow for more efficient and accurate translation (reviewed in Akashi [2001]). Our objective here is to use extensive genomic sequence data available for the genus *Saccharomyces* (Kellis et al. 2003) to provide a simple adjustment of *dS* that corrects for selection on synonymous mutations, and thereby recovers an evolutionary distance that more accurately reflects the rate of neutral evolution.

Methods

Nucleotide sequences for 3,392 orthologous open reading frames (ORFs) in *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, and *S. bayanus* were reported by Kellis et al. (2003). In our analysis, only ORFs that were represented in all four species were retained. To remove sequences with authentic or spurious frameshift mutations, which are inconsistent with the model of sequence change underlying maximum-likelihood estimation of *dS*, all sequences were translated and realigned with ClustalW version 1.8 (Thompson, Higgins, and Gibson 1994) and were then subjected to the following sliding-window filter: an ORF was eliminated from consideration if any single sequence disagreed with the majority-rule consensus protein sequence at five consecutive sites for which the consensus was defined. Visual inspection confirmed that this filter worked well to exclude putative frameshifts. A maximum-likelihood phylogeny based on nucleotide sequences was built for each ORF using PHYLIP version 3.6 (Felsenstein 2003); any ORF that did not exhibit the consensus phylogeny ([*S. cerevisiae*, *S. paradoxus*], [*S. mikatae*, *S. bayanus*]) was dropped. Synonymous divergence, *dS*, was then estimated using PAML version 3.12 (Yang 2002), with nine free parameters used to account for codon frequencies (F3×4), as in Dunn, Bielawski, and Yang (2001). Use of 60 free parameters to estimate each codon frequency individually (F61) did not significantly alter the results (for instance, Pearson's correlation coefficient between codon bias and F61 *dS* was −0.50, very similar to the correlation of −0.58 between codon bias and F3×4 *dS* [see *Results*]). Values shown here were estimated under model 0 (one *dN*/*dS* ratio for the tree), but results were extremely similar under model 1 (branch-specific *dN*/*dS* ratios).

Codon bias, as represented by the codon adaptation index, was calculated for each gene as described (Sharp and Li 1987). The set of optimal codons was taken from the 20 most highly expressed genes in *S. cerevisiae* (Arava et al. 2003), and bias values for each gene were calculated by averaging the values for all four orthologous copies.

Results and Discussion

If all synonymous sites were under equivalent selective pressure, departure from absolute neutrality would not compromise the utility of *dS* for comparing genes to assess relative divergence times, mutation rates, or nonsynonymous evolutionary rates (e.g., Lynch and Conery 2000). However, the strength of selection appears to vary widely across synonymous sites, both because the fitness penalty of inaccurate translation varies across sites (Akashi 1994) and because the optimal rate of expression varies across genes (Sharp and Li 1987; Akashi 1994; Coghlan and Wolfe 2000; Akashi 2001). Because selection in favor of preferred codons is expected to increase observable bias and to reduce the fraction of truly neutral mutations at synonymous sites, we would expect genes of higher bias to show a reduced rate of divergence at synonymous sites. This expectation is confirmed in figure 1*a*, in which each gene's synonymous divergence in a phylogeny of four species of *Saccharomyces* is plotted against that gene's average codon bias (n = 3,036 genes, each represented by

[1] Present address: Department of Systems Biology, Harvard Medical School, Boston, Massachusetts.

Key words: synonymous sites, codon bias, evolutionary genomics, dS, dN/dS, selection.

E-mail: hunter@ocf.berkeley.edu.

Mol. Biol. Evol. 22(1):174–177. 2005
doi:10.1093/molbev/msh265
Advance Access publication September 15, 2004

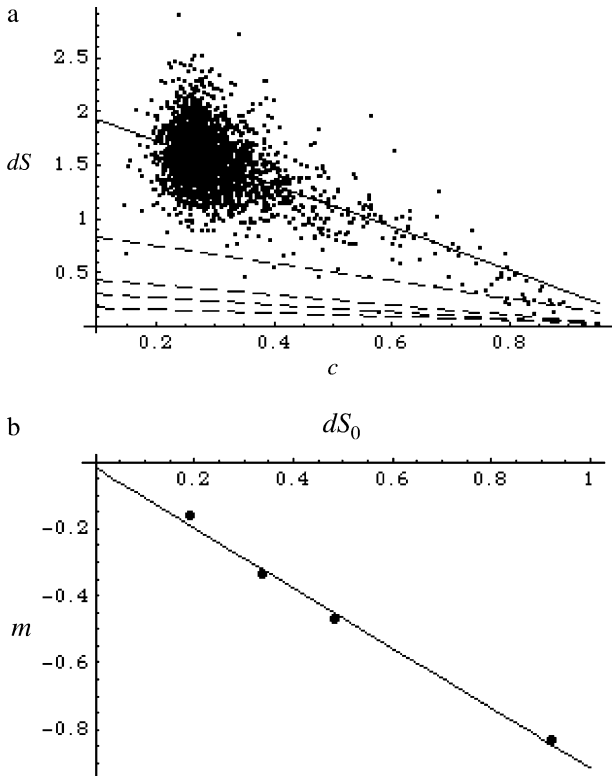

FIG. 1.—(*a*) The relationship between synonymous evolutionary divergence, *dS*, and codon bias, *c*. Points represent *dS* values summed over the complete, four-species phylogeny ([*S. cerevisiae*, *S. paradoxus*], [*S. mikatae*, *S. bayanus*]), plotted against the codon adaptation index averaged across all four species (n = 3,036, Pearson correlation coefficient $r = -0.58$, $P < 10^{-269}$). The solid line represents the best fit to these points ($dS = -2.02c + 2.13$). The dashed lines represent best-fit relationships between *dS* and *c* for each of the four yeast species individually. The lines, in order of decreasing *dS*-axis intercept, represent: *S. bayanus*, *S. mikatae*, *S. cerevisiae*, and *S. paradoxus*. Under the linear model relating *dS* to *c*, the slope of these lines, which estimates k_1t (see text) is expected to decrease in proportion to increases in their intercept, which estimates the neutral divergence, r_0t. As shown in (*b*), this is indeed the case (slope, $k_2 = -0.90$, $dS_0 = -0.016$; $r = 0.99$).

a point in figure 1*a*; Pearson's correlation coefficient $r = -0.58$; $P < 10^{-269}$). The variance in *dS* was independent of each gene's codon bias (not shown), consistent with the assumptions inherent in linear regression.

The relationship shown in figure 1*a* is approximately linear (best fit line $r^2 = 0.33$; best fit quadratic function $r^2 = 0.34$), suggesting that the rate of synonymous evolution is reduced by an amount that is linearly proportional to the level of codon bias. We, therefore, adopt the simple linear model, $dS = (r_0 + k_1c)t$, where *c* is codon bias, k_1 ($k_1 < 0$) is a constant, *t* is the total time of divergence, and r_0 is the rate of neutral evolution—that is, the rate of evolution at synonymous sites exhibiting zero bias. The quantities k_1t and r_0t are estimated by the slope (*m*) and intercept (dS_0) of the best-fit line relating *dS* to *c* for a given total divergence time *t*. For the complete tree of four yeast species, $m = -2.02$ and $dS_0 = 2.14$ (fig. 1*a*, solid line). If the linear model is appropriate, the slope *m* of the relationship between bias and divergence should decrease from zero in proportion to increases in neutral divergence dS_0, because both of these quantities estimate parameters that are proportional to the total divergence time *t*. That is, starting from slope $k_1t = 0$ at divergence time $t = 0$, the line relating bias to divergence is expected to rotate clockwise around the fixed point $c = 1$, $dS = 0$ because a sequence with perfect codon bias undergoes zero neutral divergence. We test this prediction by obtaining the best-fit line relating *dS* to *c* for each of the independent terminal branches of the four-species tree. These lines are shown in figure 1*a* (dashed lines). The yeast species they represent are, in order of decreasing dS_0, *S. bayanus*, *S. mikatae*, *S. cerevisiae*, and *S. paradoxus*. In figure 1*b*, the slope *m* of each species' line, estimating k_1t, is plotted against its intercept dS_0, estimating r_0t. As expected under the simple linear model, *m* does indeed decrease in direct proportion to neutral divergence (best-fit line in figure 1*b*: slope = −0.90; $r = 0.99$). In addition, the intercept of this relationship is not significantly different from 0 ($dS_0 = 0.016$; $P = 0.62$), in accordance with the fact that the line relating *dS* to *c* begins decreasing its slope from $k_1t = 0$ at time $t = 0$.

There are two distinct contexts in which one may require an adjustment of *dS* to take into account selection on synonymous sites. In the first context, *dS* or *dN*/*dS* values are compared across genes that share a common divergence time. For instance, to examine functional correlates of evolutionary constraint (e.g., Hirsh and Fraser 2001; Fraser et al. 2002), one would compare *dN*/*dS* among the 3,036 genes represented by points in figure 1*a*; all of these genes share the time of divergence represented by the complete, four-species phylogeny. In this context, the observed synonymous divergence of gene *i*, dS_i, the observed codon bias of gene *i*, c_i, and the best-fit estimate of k_1t (given by *m*) may be used to obtain an adjusted synonymous divergence for gene *i*, $dS_i' = dS_i - m\,c_i$. This adjustment results in a significant reduction in the variance of *dN*/*dS* (n = 3,036; *dN*/*dS* $s^2 = 0.0046$; *dN*/*dS*′ $s^2 = 0.0028$; $F = 0.60$, $P < 10^{-44}$) and significantly affects the "rank evolutionary constraint," as estimated by rank *dN*/*dS*, for a large number of genes (change in rank *dN*/*dS* > 50 for 1,553 genes). Table 1 lists the 10 yeast genes for which adjustment of *dS* has the greatest impact on rank *dN*/*dS*; that is, these are the 10 genes for which (rank *dN*/*dS*) − (rank *dN*/*dS*′) is greatest. For most of these genes, the relatively relaxed selective pressure indicated by the uncorrected *dN*/*dS* rank is at odds with the encoded protein's biological function. For instance, it seems unlikely that FBA1, which encodes fructose-bisphosphate aldolase, an abundant protein central to glycolysis, should be only the 2875th most constrained gene out of 3,036. Upon correction for codon bias, this gene's *dN*/*dS* rank improves to 333.

In another context, one may compare *dS* or *dN*/*dS* values among genes that do not share the same divergence time. For instance, in comparisons of *dN*/*dS* among duplicated genes (Lynch and Conery 2000), many paralogs do not share a common date of divergence. In this case, a single best-fit line relating *dS* to *c* cannot be used to provide an adjustment of *dS*, simply because each divergence time corresponds to a line of different slope, k_1t. However, the relationship shown in figure 1*b*, between r_0t and k_1t, provides the requisite missing information. If

Table 1
Genes with largest change in *dN/dS* rank upon adjustment of *dS*

Gene	dN/dS	dN/dS′	dN/dS rank	dN/dS′ rank	CAI	CAI rank	molecular activity	process or component
FBA1	0.225	0.023	2875	333	0.885	3035	fructose-bisphosphate aldolase	gluconeogenesis
RPS15	0.173	0.019	2607	268	0.797	3020	ribosome structural constituent	small ribosomal subunit
RPL20A	0.142	0.025	2318	394	0.742	3008	ribosome structural constituent	large ribosomal subunit
RPL9A	0.148	0.031	2379	533	0.799	3021	ribosome structural constituent	large ribosomal subunit
TEF1	0.112	0.008	1891	69	0.874	3034	translation elongation factor	ribosome
CWP2	0.144	0.037	2334	665	0.795	3018	cell wall structural constituent	cell wall
RPL3	0.103	0.014	1702	157	0.825	3025	ribosomal assembly and maint.	ribosome
PGK1	0.100	0.011	1646	112	0.832	3027	phosphoglycerate kinase	gluconeogenesis
SSB1	0.174	0.051	2617	1100	0.791	3016	ATPase and chaperone	ribosome
CDC19	0.094	0.010	1530	97	0.892	3036	pyruvate kinase	glycolysis

dN/dS values from the complete, four-species tree were estimated for 3036 genes, and adjusted as described in the text, yielding *dN/dS′*. The lowest *dN/dS* value, indicating strong constraint, corresponds to *dN/dS* rank = 1. Molecular activity, biological process, and cellular component information are from references cited by ORF on SGD (Dolinski et al. 2003).

we let k_2 ($k_2 < 0$) be the slope of the line in figure 1*b*, then $k_1 t = k_2 r_0 t$; using this equation and $dS = (r_0 + k_1 c)\, t$, we eliminate $r_0 t$ and solve for the unknown $k_1 t = k_2 dS/(1+k_2 c)$. Substituting for $k_1 t$ in the equation for corrected synonymous divergence of gene i, $dS_i' = dS_i - k_1 t c_i$, we obtain $dS_i' = dS_i/(1 + k_2 c_i)$. This provides a straightforward adjustment for *dS*, even when genes do not share a common time of divergence.

Whether the simple method of correction that is used here for *Saccharomyces* can be applied in other groups is not yet clear; a negative linear correlation would be required between *dS* and codon bias, as we observed for yeast. A negative correlation between synonymous divergence and codon bias has been reported in *Drosophila* (Shields et al. 1988; Powell and Moriyama 1997), in gene pairs resulting from the genome duplication event in an ancestor of *S. cerevisiae* (Pal, Papp, and Hurst 2001), and in *Escherichia coli* (Sharp and Li 1987; Smith and Eyre-Walker 2001). The correlation reported in *Drosophila* was subsequently found to be insignificant when using maximum-likelihood estimates of *dS* that account for unequal usage of codons (Dunn, Bielawski, and Yang 2001). Bierne and Eyre-Walker (2003) argued that this result should be revised once again because synonymous evolutionary rate should be measured not in terms of fixations per synonymous mutation, but rather in terms of synonymous fixations per physical nucleotide site. Leaving aside further discussion of the appropriate denominator for measures of synonymous evolutionary rate, we note that, for any given method of estimating *dS*, the simple correction described here would yield a metric of synonymous substitution rates that has been adjusted for codon bias. The potential problem of unequal codon usage (Dunn, Bielawski, and Yang 2001) does not apply to the relationships shown in figure 1, because we have used maximum-likelihood distance estimates that account for codon usage (Yang 2002).

The negative correlation between *dS* and codon bias in *E. coli* was suggested to be mediated by a mutational effect: if more highly expressed genes exhibit higher bias and suffer fewer mutations, the relationship between bias and *dS* could emerge without selection on silent sites (Smith and Eyre-Walker 2001). However, work on *S. cerevisiae* (Datta and Jinks-Robertson 1995; Morey, Greene, and Jinks-Robertson 2000), as well as *E. coli* (Wright, Longacre, and Reimers 1999; Klapacz and Bhagwat 2002) and *Salmonella enterica* (Hudson, Bergthorsson, and Ochman 2003), has shown that more highly expressed genes actually suffer *higher* mutation rates. This association between transcription and mutation would cause a positive correlation between *dS* and codon bias—the opposite of the one observed here. Thus, if transcription-associated mutation affects the correction we describe here, it is to make the correction somewhat conservative.

Acknowledgment

We are grateful to Dmitri Petrov for prompting this work, to Marc Feldman for advice and support, and to Jerel Davis for comments on the manuscript. This study was supported by NIH grants GM 28016 and 28424 to Marcus W. Feldman (A.E.H. and D.P.W.), by an NSF postdoctoral fellowship in bioinformatics (D.P.W.), and by an NSF predoctoral fellowship (H.B.F.).

Literature Cited

Akashi, H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics **136**:927–935.

———. 2001. Gene expression and molecular evolution. Curr Opin Genet Dev **11**:660–666.

Arava, Y., Y. Wang, J. D. Storey, C. L. Liu, P. O. Brown, and D. Herschlag. 2003. Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA **100**:3889–3894.

Bierne, N., and A. Eyre-Walker. 2003. The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. Genetics **165**:1587–1597.

Birdsell, J. A. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol. Biol. Evol. **19**:1181–1197.

Coghlan, A., and K. H. Wolfe. 2000. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. Yeast **16**:1131–1145.

Datta, A., and S. Jinks-Robertson. 1995. Association of increased spontaneous mutation rates with high levels of transcription in yeast. Science **268**:1616–1619.

Dunn, K. A., J. P. Bielawski, and Z. Yang. 2001. Substitution rates in *Drosophila* nuclear genes: implications for translational selection. Genetics **157**:295–305.

Felsenstein, J. 2003. PHYLIP (phylogeny and inference package). Version. 3.6. Distributed by the author, Department of Genetics, University of Washington, Seattle.

Fraser, H. B., A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman. 2002. Evolutionary rate in the protein interaction network. Science **296**:750–752.

Hirsh, A. E., and H. B. Fraser. 2001. Protein dispensability and rate of evolution. Nature **411**:1046–1049.

Hudson, R. E., U. Bergthorsson, and H. Ochman. 2003. Transcription increases multiple spontaneous point mutations in *Salmonella enterica*. Nucleic Acids Res. **31**:4517–4522.

Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature **423**:241–254.

Klapacz, J., and A. S. Bhagwat. 2002. Transcription-dependent increase in multiple classes of base substitution mutations in *Escherichia coli*. J. Bacteriol. **184**:6866–6872.

Langkjaer, R. B., P. F. Cliften, M. Johnston, and J. Piskur. 2003. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. Nature **421**:848–852.

Lynch, M., and J. S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. Science **290**:1151–1155.

Morey, N. J., C. N. Greene, and S. Jinks-Robertson. 2000. Genetic analysis of transcription-associated mutation in *Saccharomyces cerevisiae*. Genetics **154**:109–120.

Nei, M. 1996. Phylogenetic analysis in molecular evolutionary genetics. Ann. Rev. Genet. **30**:371–403.

Pal, C., B. Papp, and L. D. Hurst. 2001. Highly expressed genes in yeast evolve slowly. Genetics **158**:927–931.

Powell, J. R., and E. N. Moriyama. 1997. Evolution of codon usage bias in *Drosophila*. Proc. Natl. Acad. Sci. USA **94**:7784–7790.

Sharp, P. M., and W. H. Li. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4**:222–230.

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., and A. Eyre-Walker. 2001. Nucleotide substitution rate estimation in enterobacteria: approximate and maximum-likelihood methods lead to similar conclusions. Mol. Biol. Evol. **18**:2124–2126.

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

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.

Wright, B. E., A. Longacre, and J. M. Reimers. 1999. Hypermutation in derepressed operons of *Escherichia coli* K12. Proc. Natl. Acad. Sci. USA **96**:5089–5094.

Yang, Z. 2001. Adaptive molecular evolution. Pp. 327–350 *in* D. J. Balding, ed. Handbook of statistical genetics. John Wiley & Sons.

———. 2002. Phylogenetic analysis by maximum likelihood. University College, London.

Kenneth Wolfe, Associate Editor

Accepted September 9, 2004