A Dissection of Volatility in Yeast

Nina Stoletzki, John Welch,† Joachim Hermisson,* and Adam Eyre-Walker†*

*Section of Evolutionary Biology, Department Biology II, Ludwig-Maximilians-University Munich, Planegg-Martinsried, Germany; and †Centre for the Study of Evolution, University of Sussex, Brighton, United Kingdom

It has been suggested that volatility, the proportion of mutations which change an amino acid, can be used to infer the level of natural selection acting upon a gene. This conjecture is supported by a correlation between volatility and the rate of nonsynonymous substitution (dN), or the ratio of nonsynonymous and synonymous substitution rates, in a variety of organisms. These organisms include yeast, in which the correlations are quite strong. Here we show that these correlations are a by-product of a correlation between synonymous codon bias toward translationally optimal codons and dN. Although this analysis suggests that volatility is not a good measure of the selection, we suggest that it might be possible to infer something about the level of natural selection, from a single genome sequence, using translational codon bias.

Introduction

Understanding the nature of natural selection on DNA sequences is one of the central goals of molecular evolution. Plotkin, Dushoff, and Fraser (2004) and Plotkin et al. (2004) have recently suggested that it is possible to infer the level of natural selection, both positive and negative, acting upon a gene from a single genome sequence. They suggest that this can be achieved by measuring "volatility"—volatility is the proportion of point mutations in a gene, which do not yield a stop codon, which change an amino acid. They base their method on the prediction that genes which have recently undergone amino acid substitutions should be populated by codons with high volatility (Plotkin et al. 2004). In support of their thesis they show that in both *Mycobacterium* and *Saccharomyces* species, there is a correlation between volatility and the rate of nonsynonymous substitution (dN) or the ratio of nonsynonynous and synonymous substitution rates (dN/dS). This correlation is quite strong in yeast, which suggests that volatility might be a useful measure of selection.

The idea that volatility can measure the level of selection, either positive or negative, on a gene has been criticized on a number of grounds (Dagan and Graur 2004; Friedman and Hughes 2004; Sharp 2004; Chen, Emerson, and Martin 2005; Hahn et al. 2005; Nielsen and Hubisz 2005; Zhang 2005). Much of the debate has centered around the reasons why volatility is not expected to correlate to dN and dN/dS. For example, it has been suggested that volatility is unlikely to measure selection because (1) it only depends on four or five amino acids (Dagan and Graur 2004; Sharp 2004; Chen, Emerson, and Martin 2005), (2) it has low variance (Dagan and Graur 2004), and (3) simple models of evolution fail to yield a correlation between dN/dS and volatility (Dagan and Graur 2004; Nielsen and Hubisz 2005; Zhang 2005). However, volatility is correlated to dN/dS (and dN); so much of this discussion, while interesting is slightly tangential. The crucial question is why there is a correlation.

Almost all of these critiques point out that volatility is a measure of codon usage bias. As such, the apparent correlation between volatility and dN/dS in yeast may, in fact, as Hahn et al. (2005) suggest, be due to a correlation between translational codon bias and dN/dS. Although Hahn et al. suggest that the correlation between volatility and dN/dS may be due to a correlation between translational codon bias and selective constraint they do not resolve whether this is the case. They show that a measure of translational codon bias, codon adaptation index (CAI), explains more of the variance in volatility than dN/dS in yeast, but they do not pursue the matter further. Plotkin, Dushoff, and Fraser (2005) investigate the partial correlation between dN/dS and volatility controlling for CAI and show it is significant, but they fail to give the magnitude of the effect.

It, therefore, remains very unclear if the principle correlation is between dN/dS and translational codon bias, with the correlation between dN/dS and volatility a by-product of this, or whether the principle correlation is between dN/dS and volatility. Also, it might be that both translational codon bias and volatility separately correlate to dN/dS.

To investigate the matter further, we take advantage of the fact that in yeast there is a strong correlation between dN/dS (or dN) and both translational codon bias (Pal, Papp, and Hurst 2001) and volatility (Plotkin, Dushoff, and Fraser 2004) and that in yeast some of the translational optimal codons have relatively high volatility while others have relatively low volatility (table 1). It is well established that codon bias and gene expression are correlated in yeast (see e.g., Coghlan and Wolfe 2000). So volatility per amino acid is expected to increase (Ile, Leu, and Ser) or decrease (Arg and Gly) with translational codon bias or expression level (table 1). For example, the most optimal codon in yeast for argenine is AGA, which has relatively high volatility. If the principle correlation is between dN/dS (or dN) and translational codon bias, then we expect AGA usage to be negatively correlated to dN/dS (or dN), but if the principle correlation is between dN/dS (or dN) and volatility, then we expect AGA usage to be positively correlated to dN/dS (or dN).

Our results are unequivocal; in yeast dN/dS (and dN) is negatively correlated to the use of translational optimal codons for all amino acids whose synonymous codons differ in their volatility, even in those whose optimal codons have high volatility. We further show that the correlation between dN/dS (or dN) and translational optimal codon use is universal across all amino acids, including those synonymous codons which do not differ in their volatility. The

Key words: volatility, codon bias, selection, nonsynonymous substitution rate, yeast.

E-mail: stoletzki@zi.biologie.uni-muenchen.de; a.c.eyre-walker@sussex.ac.uk.

Mol. Biol. Evol. 22(10):2022–2026. 2005
doi:10.1093/molbev/msi192
Advance Access publication June 15, 2005

Table 1
Synonymous Codon Use of Volatility-Affecting Amino Acids

	Optimal Codon	Volatility of Optimal Codon	Average Volatility Per Amino Acid[a]	Average Volatility Per Amino Acid High Expression[b]	Average Volatility Per Amino Acid Low Expression[b]
Arg	AGA	0.71	0.6855	0.7065	0.6627
	CGT	0.67			
Gly	GGT	0.67	0.6648	0.6699	0.6632
Ile	ATT	0.72	0.7733	0.72	0.7960
	ATC	0.72			
Leu	TTG	0.53	0.5223	0.5243	0.5289
Ser	TCT	0.67	0.6895	0.6738	0.6929
	TCC	0.67			

[a] Given Relative Synonymous Codon Usage values of Kliman, Naheelah, and Santiago (2003).
[b] Given Relative Synonymous Codon Usage values of Sharp and Cowe (1991).

observed correlation between dN/dS (or dN) and volatility is a by-product of the correlation between dN (or dN/dS) and translational codon bias.

Materials and Methods

We downloaded the gene alignments from the four yeast species sequenced by Kellis et al. (2003). From these we excluded all genes which were not present in all the four yeast species (*Saccharomyces cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces mikatae*, and *Saccharomyces bayanus*), which did not have start and stop codons in all species, which had premature stop codons, and which had frameshifting indels. This left 1,077 genes. This is smaller than the data set analyzed by Plotkin, Dushoff, and Fraser (2005) but has less chance of containing pseudogenes.

Plotkin, Dushoff, and Fraser (2004) suggest using a statistic, volatility *P* value, to measure volatility. This is the probability of a gene having the observed volatility given the average synonymous codon use of the genes in the genome. The volatility *P* measure of Plotkin, Dushoff, and Fraser 2004 is unlikely to be a very good statistic because it will depend to some extent on gene length (Sharp 2004) and amino acid composition (Dagan and Graur 2004; Zhang 2005)—any statistic based on probability values depends on sample size, and the variance between synonymous codons for volatility differs between amino acids. To account for these shortcomings, we calculated an alternative measure, the average volatility:

$$V_{gene} = \sum V_{aa}/n,$$

where

$$V_{aa} = \frac{\sum X_i V_i}{\sum X_i}$$

and X_i is the number of times codon i is used for the amino acid aa, V_i is the volatility of that codon, and n is the number of amino acids whose synonymous codons differ in their volatility. When considering amino acids separately we used V_{aa}, the average volatility per amino acid. Note that the volatility is only affected by five amino acids whose synonymous codons differ in their volatility—Arg, Gly, Ile, Leu, and Ser (the codons of Ile only differ when the transition:transversion ratio is different from unity). We assume that the transition:transversion ratio = 4.1, to calculate the volatilities of individual codons, as suggested by Plotkin et al. (http://volatility.cgr.harvard.edu/cgi-bin/volatility.pl). We compute Plotkin's volatility *P* values using their Web site (http://volatility.cgr.harvard.edu/cgi-bin/volatility.pl).

We measured translational codon bias per gene and per amino acid separately. To measure translational codon bias, we computed the CAI according to Sharp and Li (1987) with the corrections suggested by Bulmer (1988). We also calculated the frequency of optimal codons (F_{OP}) according to the list of optimal codons for *S. cerevisiae* given by Kliman, Naheelah, and Santiago (2003). Volatility values and codon bias statistics were calculated for the *S. cerevisiae* sequence because this is the best studied of the yeasts.

We used PAML (Yang 1997) to compute dN, dS, and dN/dS for each gene using the F3 × 4 model in which codon frequencies are estimated from the nucleotide frequencies at the three codon positions. Because a physical definition of a site is more appropriate for the measurement of the synonymous substitution rate (dS), we express dS per codon (Bierne and Eyre-Walker 2003). We performed all our analyses on both dN and dN/dS. Although dN/dS is often regarded as a better measure of the selection acting upon nonsynonymous sites, it may not be in organisms, like yeast, in which there is selection on synonymous codon use. Indeed we note that there is a strong correlation between dS per codon and codon usage bias in our data (tables 2–4).

Results

Confirming the analysis of Plotkin, Dushoff, and Fraser (2004), we found a highly significant correlation between the volatility *P* value of Plotkin, Dushoff, and Fraser 2004, or average volatility, and dN/dS (or dN) per gene (table 2). We also confirm the result of Pal, Papp, and Hurst (2001) that there is a strong correlation between measures of translational codon bias (F_{OP} and CAI) and dN/dS (or dN) per gene (table 2).

So, is the observed correlation between volatility and dN/dS (or dN) due to the correlation between translational codon bias and dN/dS (or dN) or vice versa? To answer this, we look at the five volatility-affecting amino acids individually (table 3). We only observe a positive correlation

Table 2
Spearman's Rank Correlation Coefficients Between dN, dN/dS, dS Per Codon, or dS and Volatility or Translational Codon Usage Bias for Each Gene

	F_{OP}	CAI	Average Volatility	Plotkin's *P* value[a]
dN	$r = -0.419$***	$r = -0.415$***	$r = +0.256$***	$r = -0.283$***
dN/dS	$r = -0.302$***	$r = -0.292$***	$r = +0.186$***	$r = -0.224$***
dS codon	$r = -0.349$***	$r = -0.346$***	$r = +0.202$***	$r = -0.205$***

[a] Remind, Plotkin's volatility *P* value relates inversely to volatility.
*** $P < 0.001$.

between volatility and dN/dS (or dN) for three of the amino acids (Ile, Leu, and Ser). The two amino acids which show a negative correlation between volatility and dN/dS (or dN), opposing the expectation of Plotkin, Dushoff, and Fraser 2004, are those (Arg, Gly) for which high translational codon usage (in high expression genes) leads to low volatility (see table 1). There is also no indication that volatility affects the correlation; the correlation between translational codon bias and dN/dS (or dN) is as strong for Arg and Gly, as for Ile, Leu, and Ser (table 3).

The correlation between translational codon bias and dN/dS (or dN) is very consistent across amino acids—for almost every amino acid the correlation is negative and often significant, and if it is positive, the correlation is small and nonsignificant (table 4).

Discussion

We have shown that the observed correlation between dN/dS (or dN) and volatility is an incidental correlation caused by a correlation between dN/dS (or dN) and translational codon bias—dN/dS (or dN) correlates negatively with translational codon bias and volatility, for those amino acids in which the translationally optimal codons are high in volatility. This suggests that dN/dS (or dN) is not directly correlated to volatility and that volatility is therefore not the best, or even a good, predictor of dN/dS (or dN). This is not unexpected given recent theoretical work, which suggests that volatility will only be a measure of selection under rather specific conditions (Plotkin et al. 2004).

Our results may seem surprising given that Plotkin, Dushoff, and Fraser (2005) report a significant partial correlation between volatility *P* value and dN/dS in yeast using CAI to control for translational codon bias, a result we can confirm on our smaller data set (table 5). However, volatility *P* value is not normally distributed, so the probability of the partial correlation is not necessarily accurate, and the

Table 3
Spearman's Rank Correlation Coefficients Between dN, dN/dS, and dS Per Codon and Volatility or Translational Codon Usage Bias for the Five Amino Acids Effecting Volatility

		F_{OP}	CAI	Volatility
Arg	dN	−0.4267***	−0.4784***	−0.2957***
	dN/dS	−0.3359***	−0.3673***	−0.2220***
	dS	−0.3919***	−0.3780***	−0.2311***
Gly	dN	−0.4147***	−0.4862***	−0.3745***
	dN/dS	−0.2855***	−0.3538***	−0.2633***
	dS	−0.4188***	−0.4040***	−0.2758***
Ile	dN	−0.3861***	−0.4101***	+0.3936***
	dN/dS	−0.2726***	−0.2850***	+0.2783***
	dS	−0.3557***	−0.3633***	+0.3557***
Leu	dN	−0.2847***	−0.4379***	+0.0307 NS
	dN/dS	−0.1423**	−0.2744***	+0.0534 NS
	dS	−0.3729***	−0.4623***	−0.0176 NS
Ser	dN	−0.3291***	−0.3598***	+ 0.2651***
	dN/dS	−0.2012***	−0.1481***	+0.1915***
	dS	−0.3645***	−0.3637***	+0.2014***

$P < 0.01$, *** $P < 0.001$, NS = not significant.

Table 4
Spearman's Rank Correlation Coefficients Between dN, dN/dS, and dS Per Codon and Translational Codon Usage Bias for the Individual Amino Acids Not Effecting Volatility

		F_{OP}	CAI
Ala	dN	−0.2157***	−0.2919***
	dN/dS	−0.1005**	−0.1538**
	dS	−0.2949***	−0.3098***
Asn	dN	−0.2636***	−0.2828***
	dN/dS	−0.1488**	−0.1661***
	dS	−0.2300***	−0.2450***
Asp	dN	−0.074 NS	−0.0979 NS
	dN/dS	+0.0029 NS	−0.0205 NS
	dS	−0.0589 NS	−0.0690 NS
Cys	dN	−0.236***	−0.2854***
	dN/dS	−0.1666***	−0.1915***
	dS	−0.2235***	−0.2310***
Gln	dN	−0.2570***	−0.2964***
	dN/dS	−0.1584**	−0.1603**
	dS	−0.2835***	−0.3237***
Glu	dN	−0.1535**	−0.1159***
	dN/dS	−0.0581 NS	−0.0669 NS
	dS	−0.2416***	−0.2730***
His	dN	−0.0061 NS	−0.0209 NS
	dN/dS	+0.058 NS	+0.0472 NS
	dS	−0.0528 NS	−0.0748 NS
Lys	dN	−0.2753***	−0.2881***
	dN/dS	−0.1726***	−0.1701***
	dS	−0.2248***	−0.2484***
Phe	dN	−0.1617***	−0.1975***
	dN/dS	−0.0902 NS	−0.1123*
	dS	−0.1577**	−0.1891***
Pro	dN	−0.2453***	−0.3116***
	dN/dS	−0.1418**	−0.1956***
	dS	−0.2779***	−0.3227***
Thr	dN	−0.3394***	−0.3748***
	dN/dS	−0.2245***	−0.2435***
	dS	−0.2974***	−0.3156***
Tyr	dN	−0.1607**	−0.2007***
	dN/dS	−0.08 NS	−0.1113*
	dS	−0.1069*	−0.1247*
Val	dN	−0.3041***	−0.3585***
	dN/dS	−0.1935***	−0.2401***
	dS	−0.2538***	−0.3028***

* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, NS = not significant.

Table 5
Partial Correlations of Measures of Translational Codon Bias and Measures of Volatility Measures for Each Gene with dN and dN/dS

	Partial Correlations			
Control for	dN	dN/dS	dN	dN/dS
Volatility measures				
	CAI		F_{OP}	
Plotkin's *P* value	−0.273***	−0.112***	−0.228***	−0.06*
Average volatility	−0.330***	−0.179***	−0.28***	−0.122***
Translational codon bias				
	Plotkin's *P* value		Average Volatility	
CAI	−0.106***	−0.121***	−0.005 NS	−0.03 NS
F_{OP}	−0.145***	−0.150***	−0.015 NS	−0.012 NS

* $P < 0.05$, *** $P < 0.001$, NS = not significant.

significance of the partial correlation depends critically on the volatility statistic used. If we use our average volatility instead of the volatility of Plotkin et al., which will depend to some extent on gene length and amino acid composition (see *Materials and Methods*), then the partial correlation between dN/dS (or dN) and average volatility, controlling for translational codon bias, becomes very small and nonsignificant, while the partial correlation between dN/dS (or dN) and translational codon bias remains (table 5). The strongest correlations, either simple or partial, that we observe are between translational codon bias and dN/dS (or dN), which suggests that these are the primary correlations (tables 2 and 3).

It is also interesting to note that the correlation between codon bias and dN is consistently stronger than the correlation between codon bias and dN/dS. This is probably due to the fact that dS is correlated to codon bias and that this correlation is due to selection on codon usage bias and not variation in the mutation rate.

Although volatility does not appear to be a good measure of selection, Plotkin, Dushoff, and Fraser (2004) may have been correct in asserting that it may be possible to infer something about dN in a gene from a single genome sequence. A negative correlation between translational codon bias and dN has now been described in three different organisms: enteric bacteria (Sharp 1991; Rocha and Danchin 2004), Drosophila (Akashi 1994; Betancourt and Presgraves 2002; Marais et al. 2004), and yeast (Pal, Papp, and Hurst 2001), and we have shown that the correlation is consistent for all amino acids in yeast. Furthermore, although the basis of this correlation is unknown and subject to much debate (Betancourt and Presgraves 2002; Marais et al. 2004), at least one of the explanations is likely to lead to the correlation being widespread. It has been suggested that the correlation between codon bias and dN arises through a correlation in the strength of selection acting upon synonymous and nonsynonymous mutations, probably as a consequence of selection for translational accuracy—important amino acid sites in a protein will be subject to strong selection to be conserved during evolution and to be accurately translated (Akashi 1994). Thus any genome, in which selection for translational accuracy is effective, should show the correlation, and it may therefore be possible to use codon bias, maybe in combination with other information, such as amino acid composition or structural data (Tourasse and Li 2000), to predict which genes are likely to be fast-evolving genes. So, although volatility has come in for much criticism, Plotkin and colleagues may have drawn our attention to an approach to an important problem of some utility.

Acknowledgments

We thank Daniel Jeffares for some initial work, Stephan Hutter and Pieter van Beek for help with Perl, and an anonymous referee for helpful comments.

Literature Cited

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

Betancourt, A., and D. Presgraves. 2002. Linkage limits the power of natural selection in Drosophila. Proc. Natl. Acad. Sci. USA **99**(21):13616–13620.

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 synonymous substitution rate and codon usage bias. Genetics **165**:1587–1597.

Bulmer, M. 1988. Are codon usage patterns in unicellular organisms determined by selection mutation balance? J. Evol. Biol. **1**:15–26.

Chen, W., J. J. Emerson, and T. M. Martin. 2005. Not detecting selection using a single genome. Nature **433**:E6–E7.

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

Dagan, T., and D. Graur. 2004. The comparative method rules! Codon volatility cannot detect positive Darwinian selection using a single genome sequence. Mol. Biol. Evol. **22**:1260–1272.

Friedman, R., and A. L. Hughes. 2004. Codon volatility as an indicator of positive selection: data from eukaryotic genome comparisons. Mol. Biol. Evol. **22**:542–546.

Hahn, M., J. G. Mezey, D. J. Begun, J. H. Gillespie, A. D. Kern, C. H. Langley, and L. Moyle. 2005. Codon bias and selection on single genomes. Nature **433**:E5.

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

Kliman, R. M., I. Naheelah, and M. Santiago. 2003. Selection conflicts, gene expression, and codon usage trends in yeast. J. Mol. Evol. **57**:98–109.

Marais, G., T. Domazet-Loso, D. Tautz, and B. Charlesworth. 2004. Correlated evolution of synonymous and nonsynonymous sites in *Drosophila*. J. Mol. Evol. **59**:771–779.

Nielsen, R., and M. J. Hubisz. 2005. Detecting selection needs comparative data. Nature **433**:E6.

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

Plotkin, J. B., J. Dushoff, M. M. Desai, and H. B. Fraser. 2004. Synonymous codon usage and selection on proteins. (http://www.arxiv.org/abs/q-bio/0410013).

Plotkin, J. B., J. Dushoff, and H. B. Fraser. 2004. Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*. Nature **428**:942–945.

———. 2005. Reply. Nature **433**:E7–E8.

Rocha, E. P. C., and A. Danchin. 2004. An analysis of determinants of amino acid substitution rates in bacterial proteins. Mol. Biol. Evol. **21**(1):108–116.

Sharp, P. M., 1991. Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. J. Mol. Evol. **33**:23–33.

Sharp, P. M. 2004. Gene "volatility" is most unlikely to reveal adaptation. Mol. Biol. Evol. **22**:807–809.

Sharp, P. M., and E. Cowe. 1991. Synonymous codon usage in Saccharomyces cerevisiae. Yeast **7**:657–678.

Sharp, P. M., and W.-H. Li. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15**:1281–1295.

Tourasse, N., and W.-H. Li. 2000. Selective constraints, amino acid composition and the rate of protein evolution. Mol. Biol. Evol. **17**(4):656–664.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**:555–556.

Zhang, J. 2005. On the evolution of codon volatility. Genetics **169**:495–501.

William Martin, Associate Editor

Accepted June 9, 2005