LETTERS

Estimating Selection Pressures from Limited Comparative Data

Joshua B. Plotkin, Jonathan Dushoff,† Michael M. Desai,‡ and Hunter B. Fraser§*

*Department of Biology, University of Pennsylvania; †Department of Ecology and Evolutionary Biology, Princeton University; ‡Department of Molecular and Cellular Biology and Department of Physics, Harvard University; and §Broad Institute of Harvard and the Massachusetts Institute of Technology, Cambridge, Massachusetts

We recently introduced a novel method for estimating selection pressures on proteins, termed "volatility," which requires only a single genome sequence. Some criticisms that have been levied against this approach are valid, but many others are based on misconceptions of volatility, or they apply equally to comparative methods of estimating selection. Here, we introduce a simple regression technique for estimating selection pressures on all proteins in a genome, on the basis of limited comparative data. The regression technique does not depend on an underlying population-genetic mechanism. This new approach to estimating selection across a genome should be more powerful and more widely applicable than volatility itself.

The volatility of a codon is defined as the proportion of its nontermination point–mutational neighbors that encode a different amino acid (Plotkin et al. 2004). The volatility *P* value of a gene quantifies the degree to which the gene's total (or "raw") volatility is significantly elevated or depressed compared with the codon usage in the genome as a whole, controlling for the gene's amino acid sequence (Plotkin and Dushoff 2003; Plotkin et al. 2004). The *P* value is computed by comparing the observed gene sequence to many random sequences that are identical at the amino acid level but whose codons are drawn according to the genome-wide codon usage (Plotkin et al. 2004). The *P* value is 2-sided in an atypical sense: *P* near 0 indicates the gene has significantly elevated volatility, and *P* near 1 indicates the gene has significantly depressed volatility. We have proposed that the volatility *P* values of genes (not raw volatilities) reflect the relative selection pressures experienced by proteins in a genome (Plotkin and Dushoff 2003; Plotkin et al. 2004).

We have previously demonstrated that volatility *P* values are significantly correlated with traditional estimates of selection pressures on proteins, significantly depressed among surface antigens of pathogens known to experience positive selection, and significantly elevated among the genes conserved between bacterial species and the genes essential for bacterial viability (Plotkin and Dushoff 2003; Plotkin et al. 2004).

In addition, the antigens of several other pathogens also exhibit significantly elevated volatility. In *Borrelia burgdorferi*, which causes Lyme disease, most proteins have unknown function, but the proteins P35 and P37 have been identified as immunogenic (Fikrig et al. 1997). We might therefore expect that these proteins experience positive selection driven by the host immune system. Indeed, the 3 proteins that are annotated as P37 antigens and lack frameshift mutations are significantly biased toward elevated volatility: 2 of the 3 are among the 20 lowest volatility *P* values in the entire genome, which is a significant enrichment (hypergeometric $P < 7 \times 10^{-6}$). Similarly, in *Yersinia pestis*, the causative agent of plague, recent work has identified 24 hypervariable, virulence-related genes present in multiple-sampled isolates (Hinchliffe et al. 2003). Again, these genes are significantly enriched for low volatility *P* values: the 4 most volatile genes in the genome, and 5 of the 10 most volatile genes, all belong to the list of putative virulence factors (hypergeometric $P < 2 \times 10^{-9}$). Despite these highly significant results, we stress that not all antigens exhibit elevated volatility. This is likely due both to a lack of strong positive selection on some antigens and to the limited power of volatility to detect selection.

Many criticisms of volatility have arisen from simple misunderstandings of the method. Several authors have suggested that empirical results using volatility are artifacts caused by the length or amino acid composition of rapidly evolving proteins (Dagan and Graur 2005; Friedman and Hughes 2005; Nielsen and Hubisz 2005; Sharp 2005; Stoletzki et al. 2005). These suggestions are clearly incorrect because our empirical results are all based on volatility *P* values that control exactly for each gene's amino acid sequence (Plotkin et al. 2004). A gene with more informative sites can achieve a more extreme *P* value (as with any statistical test of selection, more data give more power), but a gene's amino acid content or length cannot possibly bias its volatility *P* value toward 0 or 1. This misunderstanding may have arisen because others have mistakenly analyzed raw volatility instead of volatility *P* values (Dagan and Graur 2005; Friedman and Hughes 2005; Sharp 2005). Additionally, simulations that fail to account for population variability (Dagan and Graur 2005; Nielsen and Hubisz 2005; Zhang 2005) do not find any effect of selection on volatility (Plotkin et al. 2005), whereas more realistic simulations that properly account for population variability find significant effects of selection on volatility (Golding and Strobeck 1982; Archetti 2006; Plotkin et al. forthcoming).

There remain, however, many practical limitations of the volatility method. The power to detect negative selection depends strongly on the product of the effective population size and mutation rate (Chen et al. 2005), and so volatility is applicable only to some viral and microbial species (Plotkin et al. forthcoming). Even when this product is large, many sites are required to detect selection (Plotkin et al. forthcoming). In addition, we and others

Key words: volatility, codon usage, selection, comparative method.

E-mail: jplotkin@sas.upenn.edu.

Mol. Biol. Evol. 23(8):1457–1459. 2006
doi:10.1093/molbev/msl021
Advance Access publication June 5, 2006

have pointed out that differential selection on synonymous sites—for example, selection for translational optimality that varies across the genome (Akashi and Eyre-Walker 1998)—will distort estimates of selection on proteins based on volatility (Plotkin et al. 2004; Hahn et al. 2005; Stoletzki et al. 2005), but it should be noted that such processes will likewise distort estimates based on homologous sequence comparison (Sharp and Li 1987; Hirsh et al. 2005; Chamary et al. 2006).

Given the limitations of volatility, we have developed an alternative method to estimate selection pressures on all proteins in a sequenced genome, using only limited comparative data. This method is designed to approximate dN/dS values (Goldman and Yang 1994) on the basis of synonymous codon usage (Stoletzki et al. 2005). Starting from a subset of genes with known orthologs and measured dN/dS values, we first regress synonymous codon usage against dN/dS, and we then extrapolate dN/dS values for the remaining genes in the genome on the basis of their codon usage. An example of this technique is given in table 1, which shows the best-fit linear combination of codon usage that predicts dN/dS (after the square root transformation, to improve normality) for 2952 genes in *Saccharomyces cerevisiae*. As this analysis demonstrates, synonymous codon usage alone explains a large amount of the variation in dN/dS ($r = 0.63$), even after dS has been corrected for selection on silent sites (Hirsh et al. 2005). The same technique applied to *Escherichia coli* also yields a linear combination of synonymous codon usage that is predictive of dN/dS ($r = 0.54$ for 1849 *E. coli* genes with orthologs in *Vibrio cholerae*). The technique also works in *Drosophila melanogaster* ($r = 0.52$ for 11 700 genes with orthologs in *Drosophila pseudoobscura*) as well as *Homo sapiens* ($r = 0.43$ for 11 848 genes with orthologs in *Mus musculus*).

There is the potential concern that a regression for dN/dS calibrated on a subset of a genome may not yield accurate estimates for the remainder of the genome—especially considering that genes with identifiable orthologs comprise a biased subset of slowly evolving proteins. To address this concern, we have repeated our analysis of *S. cerevisiae* by regressing codon usage against dN/dS on only those 1350 genes with orthologs in the distant species *Candida albicans*. The resulting best-fit linear combination of codon usage is still a good predictor of dN/dS for the remaining 1602 genes with orthologs only in more closely related species ($r = 0.61$), despite the fact that these genes differ qualitatively from the genes used in the regression. Other characteristics of genes may be incorporated as independent variables in such regressions in order to improve their predictive power for dN/dS. For example, including each gene's amino acid frequencies improves predictive power by up to 50%.

The method introduced here is not the same as volatility, but it substantiates the same underlying principle: synonymous codon usage contains information about selection pressures on proteins, and it may be used to estimate selection on proteins that cannot be studied through comparative analysis. Our simple regression method does not specify a mechanism or depend on an underlying population-genetic model, and it is therefore free of some criticisms that might apply to volatility. Although this method, like volatility itself, provides less precise estimates of selection than homologous sequence comparison, the method requires fewer data, and it allows one to screen an entire genome for candidate proteins under strong selection.

Table 1
A Regression of Synonymous Codon Usage against $\sqrt{dN/dS}$ for 2952 *Saccharomyces cerevisiae* Genes

Codon	Coefficients
TGC	0.0092
TTC	−0.0096
CTC	0.2187**
CTG	−0.0315
CTA	0.0089
CTT	0.0945*
TTG	0.0100
CCC	−0.0088
CCG	0.0149
CCA	0.0225
CAG	0.0038
TGA	−0.0043
TAG	−0.0047
GCC	0.0185
GCG	0.0447*
GCA	0.0466*
GAC	0.0023
CAC	0.0153
CGC	0.1183**
CGG	0.1832**
CGA	0.1421**
CGT	−0.0306
AGG	0.0432*
ACC	−0.0219
ACG	0.0139
ACA	−0.0134
TAC	0.0020
ATC	−0.0044
ATA	0.0901**
GGC	0.0797**
GGG	0.0322
GGA	0.1142**
GAG	−0.0025
AAG	−0.0416*
AGC	0.0995**
AGT	0.0289
TCC	0.0120
TCG	0.0014
TCA	0.0317
GTC	−0.0369
GTG	−0.0423*
GTA	0.0276
AAC	−0.0157
Constant	0.1125**

NOTE.—The sum of these coefficients times the relative frequency of each codon (compared with its synonyms) is highly predictive of $\sqrt{dN/dS}$ ($r = 0.63$). Asterisks indicate coefficients that deviate significantly from 0 (*$P < 0.01$; **$P < 0.000\,001$).

Methods

For the purpose of calculating volatility values, we estimated the median transition/transversion biases for the genomes of *S. cerevisiae*, *B. burgdorferi*, and *Y. pestis* as $\kappa = 4.1$, $\kappa = 1.3$, and $\kappa = 2.0$, respectively, using the method of Yang (1997). Orthology assignment and dN/dS values for *H. sapiens* were obtained from the ENSEMBL database (Birney et al. 2006). For *D. melanogaster*, orthologs were assigned by reciprocal best Blast, and dN/dS values were calculated according to Yang and Nielsen (2000). For *S. cerevisiae*,

orthologs were assigned by Kellis et al. (2003). The dN/dS values were corrected for selection on synonymous sites (Hirsh et al. 2004), although similar results are obtained without such correction. Orthologs in *C. albicans* were identified according to Wall et al. (2003). All data sets are available on request.

Acknowledgments

The authors are grateful for discussions with Daniel Fisher. J.D. acknowledges support from NIH-P50-GM071508 to D. Botstein. J.B.P. acknowledges support from the Burroughs Wellcome Fund and the W. F. Milton Fund.

Literature Cited

Akashi HA, Eyre-Walker A. 1998. Translational selection and molecular evolution. Curr Opin Genet Dev 8:688–93.

Archetti M. 2006. Genetic robustness and selection at the protein level for synonymous codons. J Evol Biol 19:353–65.

Birney E, Andrews D, Caccamo M, et al. (50 co-authors). 2006. Ensembl 2006. Nucleic Acids Res 34:D556–61.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: nonneutral evolution at synonymous sites in mammals. Nat Rev Genet 7:98–108.

Chen Y, Emerson JJ, Martin TM. 2005. Codon volatility does not detect selection. Nature 433:E6–7.

Dagan T, Graur D. 2005. The comparative method rules! Codon volatility cannot detect positive Darwinian selection using a single genome sequence. Mol Biol Evol 22:496–500.

Fikrig E, Barthold SW, Sun W, Feng W, Telford SRI, Flavell RA. 1997. Borrelia burgdorferi P35 and P37 proteins, expressed in vivo, elicit protective immunity. Immunity 6:531–9.

Friedman R, Hughes AL. 2005. Codon volatility as an indicator of positive selection: data from eukaryotic genome comparisons. Mol Biol Evol 22:542–6.

Golding GB, Strobeck C. 1982. Expected frequencies of codon use as a function of mutation rates and codon fitnesses. J Mol Evol 18:379–86.

Goldman N, Yang Z. 1994. Codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–36.

Hahn MW, Mezey JG, Begun DJ, Gillespie JH, Kern AD, Langley CH, Moyle LC. 2005. Codon bias and selection on single genomes. Nature 433:E5–6.

Hinchliffe SJ, Isherwood KE, Stabler RA, Prentice MB, Rakin A, Nichols RA, Oyston PC, Hinds J, Titball RW, Wren BW. 2003. Application of DNA microarrays to study the evolutionary genomics of *Yersinia pestis* and *Yersinia pseudotuberculosis*. Genome Res 13:2018–29.

Hirsh AE, Fraser HB, Wall DP. 2005. Adjusting for selection on synonymous sites in estimates of evolutionary distance. Mol Biol Evol 22:174–77.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–54.

Nielsen R, Hubisz MJ. 2005. Detecting selection needs comparative data. Nature 433:E6.

Plotkin JB, Dushoff J. 2003. Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. Proc Natl Acad Sci USA 100:7152–7.

Plotkin JB, Dushoff J, Desai MM, Fraser HB. Codon usage and selection pressures on proteins. J Mol Evol. Forthcoming.

Plotkin JB, Dushoff J, Fraser HB. 2004. Detecting selection using a single genome sequence of M. tuberculosis and P. falciparum. Nature 428:942–5.

Plotkin JB, Dushoff J, Fraser HB. 2005. Communication arising: evolutionary genomics. Nature 433:E7–8.

Sharp PM. 2005. Gene "volatility" is most unlikely to reveal adaptation. Mol Biol Evol 22:807–9.

Sharp PM, Li WH. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol 4:222–30.

Stoletzki N, Welch J, Hermisson J, Eyre-Walker A. 2005. A dissection of volatility in yeast. Mol Biol Evol 22:2022–6.

Wall DP, Fraser HB, Hirsh AE. 2003. Detecting putative orthologs. Bioinformatics 19:1710–1.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–6.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17:32–43.

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

William Martin, Associate Editor

Accepted April 28, 2006