On the Varied Pattern of Evolution of 2 Fungal Genomes: A Critique of Hughes and Friedman

Ziheng Yang

Department of Biology, University College London, London, United Kingdom

A number of statistical tests have been proposed to detect positive Darwinian selection affecting a few amino acid sites in a protein, exemplified by an excess of nonsynonymous nucleotide substitutions. These tests are often more powerful than pairwise sequence comparison, which averages synonymous (d_S) and nonsynonymous (d_N) rates over the whole gene. In a recent study, however, Hughes AL and Friedman R (2005. Variation in the pattern of synonymous and nonsynonymous difference between two fungal genomes. *Mol Bio Evol.* 22: 1320–1324) argue that d_S and d_N are expected to fluctuate along the sequence by chance and that an excess of nonsynonymous differences in individual codons is no evidence for positive selection. The authors compared codons in protein-coding genes from the genomes of 2 yeast species, *Saccharomyces cerevisiae* and *Saccharomyces paradoxus*. They calculated the proportions of synonymous and nonsynonymous differences per site (p_S and p_N) in every codon and discovered that p_N is often greater than p_S and that among some codons p_S and p_N are negatively correlated. The authors argued that these results invalidate previous tests of codons under positive selection. Here I discuss several errors of statistics in the analysis of Hughes and Friedman, including confusion of statistics with parameters, arbitrary data filtering, and derivation of hypotheses from data. I also apply likelihood ratio tests of positive selection to the yeast data and illustrate empirically that Hughes and Friedman's criticisms on such tests are not valid.

Introduction

Recently, several statistical tests have been introduced to compare protein-coding DNA sequences across species to detect positive Darwinian selection affecting only a few amino acid sites in the protein. These methods rely on the rationale that if nonsynonymous (amino acid altering) mutations offer a fitness advantage, they will be driven to fixation by positive selection, resulting in higher nonsynonymous (d_N) than synonymous (d_S) substitution rates. A simple approach exploiting this idea is to reconstruct ancestral sequences on the phylogeny and to count synonymous and nonsynonymous changes at each codon along the tree to test whether $d_N > d_S$ at each codon (Fitch et al. 1997; Suzuki and Gojobori 1999). Another heuristic approach is to use a sliding window along the sequence alignment and test for $d_N > d_S$ in each window. A more rigorous approach is to construct a likelihood ratio test (LRT) to compare 2 nested codon substitution models of variable ω ($=d_N/d_S$) among sites, one of which allows for sites with $\omega > 1$ and the other of which does not (Nielsen and Yang 1998; Yang et al. 2000). By allowing for variable selective pressures among amino acid sites, those methods have in general more power to detect positive selection than the early approach of averaging d_N and d_S along the whole protein (e.g., Anisimova et al. 2001).

Recently, Hughes and Friedman (2005), referred to later as "HF05," argue that such methods are invalid. The authors compared codons in the protein-coding genes from the complete genomes of *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* and calculated the proportions of synonymous nucleotide differences per synonymous site and of nonsynonymous differences per nonsynonymous site (p_S and p_N) in every codon. They discovered a negative correlation between p_S and p_N and argued that a higher p_N than p_S (or a higher d_N than d_S) in some codons may occur by chance and does not imply positive selection.

Here, I counter the analysis of HF05 on statistical grounds through a reanalysis of the yeast data. I offer an explanation for the negative correlation between p_S and p_N observed in HF05, which makes it clear that the correlation has no biological significance. In addition, I apply LRTs (Nielsen and Yang 1998; Yang et al. 2000) to the yeast data and argue that the results of such analysis are sensible biologically.

Issues in the Analysis of Hughes and Friedman

HF05 concatenated the 4,133 protein-coding genes from *S. cerevisiae* and *S. paradoxus* into one "supergene" and then conducted a codon-by-codon analysis. The Nei and Gojobori (1986, NG86) method was used to estimate the numbers of synonymous and nonsynonymous sites (s and n) and the numbers of synonymous and nonsynonymous differences (s_d and n_d) in every codon. The proportions of synonymous and nonsynonymous differences per nucleotide site were then calculated as $p_S = s_d/s$ and $p_N = n_d/n$. The authors conducted various tests using p_S and p_N. For example, they found an excess of codons in which $p_N > p_S$ relative to "random expectation," which the authors equate with "random pairing of observed p_S and p_N values." Furthermore, by removing codons that are identical between the 2 species, HF05 discovered a negative correlation between p_S and p_N.

Let us consider an idealized case. Imagine a "regular" genetic code in which every codon is 4-fold degenerate; that is, 16 amino acids are encoded by 64 sense codons, and the first- and second-codon positions completely determine the amino acid. The numbers of synonymous and nonsynonymous sites in every codon are then $s = 1$ and $n = 2$, according to the NG86 procedure. Suppose that we filter the data even further and use only codons that differ at one position between the 2 species. As the single difference must be either synonymous or nonsynonymous, one of s_d and n_d must be 0 and the other must be 1, with $s_d + n_d = 1$.

Key words: genome evolution, likelihood ratio test, nonsynonymous substitution, positive selection, synonymous substitution.

E-mail: z.yang@ucl.ac.uk.

Mol. Biol. Evol. 23(12):2279–2282. 2006
doi:10.1093/molbev/msl122
Advance Access publication September 18, 2006

Table 1
Log-likelihood Values and Parameter Estimates under Various Site Models for the Concatenated Data

Model	p	ℓ	Estimates of Parameters
M0 (one ratio)	1	−10,419,399.84	$\hat{\omega}=0.111$
M1a (neutral)	2	−10,407,049.69	$\hat{p}_0=0.918$ ($\hat{p}_1=0.082$), $\hat{\omega}_0=0.054$ ($\omega_1=1$)
M2a (selection)	4	−10,406,700.21	$\hat{p}_0=0.925$, $\hat{p}_1=0.073$ ($\hat{p}_2=0.001$), $\hat{\omega}_0=0.059$ ($\omega_1=1$), $\hat{\omega}_2=138$
M7 (beta)	2	−10,407,311.81	$\hat{p}=0.123$, $\hat{q}=0.805$
M8 (beta&ω)	4	−10,406,754.33	$\hat{p}_0=0.997$ ($\hat{p}_1=0.003$), $\hat{p}=0.182$, $\hat{q}=1.297$, $\hat{\omega}=9.528$

NOTE.—p is the number of parameters in the ω distribution. Other parameters common to all models are the sequence divergence t, transition/transversion rate ratio κ, and the 9 parameters for the base compositions at the 3 codon positions. Estimates of t range from 0.40 to 0.57 nucleotide substitutions per codon among models, whereas those of κ range from 4.3 to 4.6.

The correlation between $p_S = s_d/s = s_d$ and $p_N = n_d/n = n_d/2$ will be −1. This perfect negative correlation clearly does not mean anything biologically. As an analogy, consider n tosses of a coin. The counts of heads and tails (x_i and y_i) in every toss i have the correlation −1 as $x_i + y_i = 1$, but this correlation has no bearing on whether or not the coin is fair or whether one fair coin or several coins with different biases are tossed. For 2 reasons, the correlation between p_S and p_N in every codon calculated in HF05 was not exactly −1. First, the real genetic code is not regular and the numbers of synonymous and nonsynonymous sites (s and n) fluctuate somewhat among codons. Second, HF05 used not only codons with 1 difference but also those with 2 or 3 differences, even though the former are by far more frequent than the latter. Nevertheless, it is clear that the negative correlation between p_S and p_N is due to the definition and calculation of those quantities and has no relevance to the validity or invalidity of the LRT of positive selection.

HF05 tested whether the number of codons in which $p_N > p_S$ significantly exceeds the "neutral expectation." Their null hypothesis for this test was not stated, but the null distribution was generated by "random pairing of p_S and p_N" calculated from the data. In the idealized case mentioned above, p_S and p_N can take only 2 sets of values: 1) $p_S = 0$ if and only if $p_N = ½$ and 2) $p_S = 1$ if and only if $p_N = 0$. Random pairing of p_S and p_N is simply impossible under any model. It has nothing to do with neutral evolution and should not be taken as a null hypothesis.

Reanalysis of the Yeast Data

HF05 criticized previous methods for detecting amino acid sites under positive selection. They discussed random fluctuations in p_S and p_N and argue that "available methods do not appear to include any effective controls for such stochastic variation." This statement reflects a misunderstanding of statistical hypothesis testing, the principal objective of which is to control for stochastic variation in the data. Consider the LRT comparing codon models M1a (neutral) with M2a (selection) (Nielsen and Yang 1998; Yang et al. 2005). M1a assumes 2 classes of sites with $0 \le \omega_0 < 1$ and $\omega_1 = 1$, respectively. M2a is a more general model and adds an extra class of sites with $\omega_2 \ge 1$. Under M1a, many codons, especially those with $\omega_1 = 1$, will show $p_N > p_S$ just by chance, but the LRT will be significant in no more than 5% of the data sets when the test is conducted at the 5% level. M1a is rejected only when M2a, by including a site class with $\omega > 1$, explains the fluctuations in the data much better than M1a can. Of course, if neither M1a nor M2a fits the data, the calculated p values may not be accurate. The robustness of such tests to violations of model assumptions is an important issue and has been studied by applying multiple codon models (such as M7 and M8; see below) to the same data and by using computer simulations (e.g., Anisimova et al. 2001; Wong et al. 2004).

Here I provide a likelihood analysis of the yeast data. All 2,053,314 codons in the 4,133 genes are used. Between the 2 species, 1,475,239 codons are identical, 523,356 have 1 nucleotide difference, 50,574 2 differences, and 4,145 3 differences. First, the concatenated supergene was analyzed under model M0 (one ratio; F3 × 4 model of codon usage) (Goldman and Yang 1994; Yang 1997). The estimates are $\hat{d}_N=0.0414$, $\hat{d}_S=0.3720$, and $\hat{\omega}=0.1112$, which are averages over the whole genome. The small estimate of ω indicates that on average the yeast proteins are under strong purifying selection.

Table 1 summarizes the analysis of the concatenated data under site models M1a (neutral), M2a (selection), M7 (beta), and M8 (beta&ω) (Nielsen and Yang 1998; Yang et al. 2000, 2005). Models 1a and M2a are described above. Model M7 (beta) assumes a beta distribution for ω, so that $0 \le \omega \le 1$ and no Darwinian selection is permitted. M8 (beta&ω) adds an extra site class with $\omega_s > 1$. The 2 LRTs that compare M1a with M2a and M7 with M8 are both significant (table 1). Many codons are included in the data set, so it is not surprising that the tests are significant. The maximum likelihood estimates under M2a suggest 0.1% of sites are under positive selection with a very large $\hat{\omega}_2=138$, whereas M8 suggests that 0.3% of sites are under positive selection with $\hat{\omega}_s=9.53$. As it is very difficult to distinguish between a slightly smaller proportion of sites under stronger positive selection and a larger proportion of sites under weaker selection, the estimates under models M2a and M8 cannot be considered to be very different. The Bayes Empirical Bayes approach (Yang et al. 2005) was used to calculate the posterior probability that every codon is from the site class of positive selection. No sites reached the 95% cutoff. Under M2a, the highest posterior probability is ~0.74, whereas under M8, 155 codons have highest posterior probabilities, in the range (0.90, 0.93). Although the LRTs provide strong evidence for presence of sites under positive selection, the data are not informative enough to allow for their reliable identification.

I then apply models M0 (one ratio), M1a (neutral), and M2a (selection) to the 4,133 genes separately. Histograms

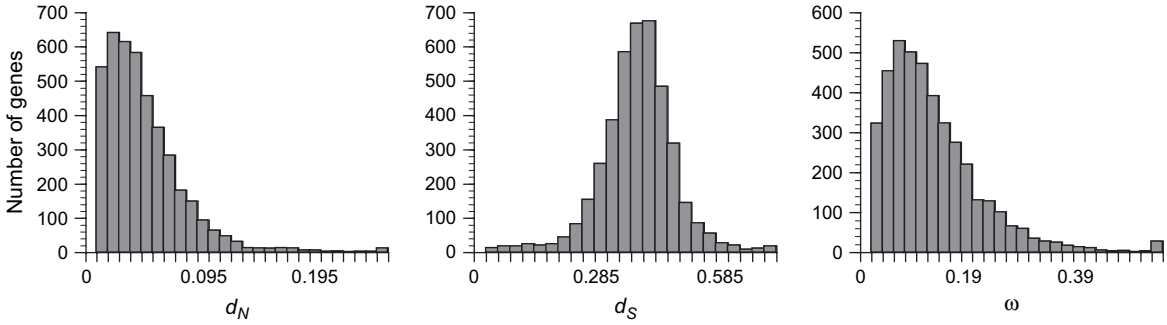

FIG. 1.—Histograms of d_N, d_S, and ω among 4,133 genes estimated under codon model M0 (one ratio). The F3 × 4 model of codon usage is assumed.

of d_N, d_S, and ω estimated under M0 are shown in figure 1. Two genes (YNL326C and YNL328C) had $\hat{\omega} = \infty$ as there is no synonymous difference between the 2 species, and 3 genes (YKL165C-A, YLR415C, and YDL240C-A) had $\hat{\omega}$ very slightly greater than 1, whereas in all other genes $\hat{\omega} < 1$. None of those ω values is significantly greater than 1 (results not shown). The approach of averaging ω over all sites thus fails to detect positive selection in any gene. I then apply the LRT comparing models M1a and M2a. In 126 genes, the test was significant at the 5% level. The site-based test is noted to be far more powerful than the test comparing d_N and d_S averaged over all codons in the gene. The 126 genes include some cell-wall proteins, proteins involved in bud-site selection, and a number of hypothetical proteins. The full list is given in table S1 as online supplementary information.

It should be noted that application of the LRT to many genes may lead to false rejections of the null hypothesis by chance, due to multiple testing. If M1a is true in every gene, we expect 4,133 × 5% = 207 genes to show significant results just by chance. Yet, only 126 genes reached significance. The false-positive rate of the test is clearly lower than 5% in those data sets, contra the claim of HF05 that the LRT, by allowing variables d_N and d_S among sites, should generate excessive false positives. In this case, the lack of power of the LRT in data sets of only 2 sequences is a more serious concern (e.g., Anisimova et al. 2001). In particular, models M1a and M2a are designed to reduce false positives even when the null model M1a is seriously violated. The insistence on a site class with $\omega_1 = 1$ makes the test rather robust to violations of assumptions. For example, in many simulation studies, the false-positive rate of the test comparing M1a with M2a was lower than the significance level (e.g., Wong et al. 2004; Yang et al. 2005). Most real genes probably do not have any codons undergoing entirely neutral evolution with $\omega = 1$. In such data sets, the LRT comparing M1a and M2a tends to have false-positive rates lower than the significance level.

It is possible to apply multiple-test corrections to the LRT applied to all genes. However, it appears biologically obvious that some genes in the yeast genome are affected by positive selection, and a formal test for the presence of such genes may not be necessary. Instead, the list in table S1 (supplementary material online) includes the top few genes that are most likely to be under positive selection from this analysis. It will be interesting to examine whether the results hold up when more genomes become available to allow a more powerful analysis.

The Role of Models in Biological Data Analysis

The discovery in HF05 of a negative correlation between p_S and p_N was apparently not motivated by any evolutionary theory but was rather a product of seeking unusual patterns in the data. Such data dredging may be used meaningfully to discover unexpected relationships to formulate a hypothesis to be tested with future data. It has to be treated with caution if the hypothesis is derived from the data and then tested using the same data, as in HF05.

The second issue in the analysis of HF05 concerns exclusion of data. HF05 removed codons with no difference between the 2 species without accounting for the fact that a nonrandom part of the observed data is removed. We note that removing outliers or mistaken data points after careful inspection may be reasonable and important. Furthermore, data censoring is sometimes unavoidable as the data may be compiled from past experiments not designed to address the hypothesis being tested. However, it alters the sampling distribution of the data and has to be taken into account in the analysis.

A third issue in HF05 is the lack of a well-specified statistical model and of a clear distinction between parameters and statistics. Parameters are unknown constants about which we would like to draw inferences. Statistics are calculated from the data and have distributions specified by the model and parameter values. Statistical hypotheses should be formulated by placing constraints on parameters in the model rather than on statistics observed in the data. This is the case whether a parametric or nonparametric approach is taken in the analysis. In this regard, p_S and p_N are statistics calculated from the data. HF05 use p_S and p_N to formulate the hypothesis that p_S and p_N have zero correlation. The authors' resistance to a clear formulation of model assumptions is further illustrated in a similar analysis of protein-coding genes from the mouse, rat, and human genomes, which makes similar mistakes (Friedman and Hughes 2005). Here the authors claim that they "use simple methods that do not depend on any model of nucleotide substitution, but rather on comparative analysis of patterns of nucleotide difference." This claim, bold as it is, is not justified. By failing to specify the model assumptions explicitly, it is not clear how statistical inferences can be

drawn from the analyses in either Hughes and Friedman (2005) or Friedman and Hughes (2005). Statistical inference requires a statistical model.

Supplementary Material

Table S1, which lists 126 genes detected to be under positive selection by the LRT comparing models M1a and M2a, is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

I thank Dr Nick Goldman and 2 anonymous referees for critical and constructive comments on an earlier version of this article. I am grateful to Drs Austin Hughes and Robert Friedman for providing the yeast data. This study is supported by a grant from the Biotechnological and Biological Sciences Research Council (United Kingdom).

Literature Cited

Anisimova M, Bielawski JP, Yang Z. 2001. The accuracy and power of likelihood ratio tests to detect positive selection at amino acid sites. Mol Biol Evol. 18:1585–1592.

Fitch WM, Bush RM, Bender CA, Cox NJ. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci USA. 94:7712–7718.

Friedman R, Hughes AL. 2005. The pattern of nucleotide difference at individual codons among mouse, rat, and human. Mol Biol Evol. 22:1285–1289.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11:725–736.

Hughes AL, Friedman R. 2005. Variation in the pattern of synonymous and nonsynonymous difference between two fungal genomes. Mol Biol Evol. 22:1320–1324.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 148:929–936.

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

Wong WSW, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics. 168:1041–1051.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. [Internet]. 13:555–556. Available from: http://abacus.gene.ucl.ac.uk/software/paml.html. Accessed on October 12, 2006.

Yang Z, Nielsen R, Goldman N, Pedersen A-MK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 155:431–449.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22:1107–1118.

William Martin, Associate Editor

Accepted September 12, 2006