Variation in the Pattern of Synonymous and Nonsynonymous Difference Between Two Fungal Genomes

Austin L. Hughes and Robert Friedman

Department of Biological Sciences, University of South Carolina

The proportion of synonymous nucleotide differences per synonymous site \( (p_S) \) and the proportion of nonsynonymous differences per nonsynonymous site \( (p_N) \) were computed at 1,993,217 individual codons in 4,133 protein-coding genes between the two yeast species \( \text{Saccharomyces cerevisiae} \) and \( \text{Saccharomyces paradoxus} \). When the modified Nei-Gojobori method was used, significantly more codons with \( p_N > p_S \) were observed than expected, based on random pairing of observed \( p_S \) and \( p_N \) values. However, this finding was most likely explained by the presence of a strong negative correlation between the number of synonymous differences and the number of nonsynonymous differences at codons with at least one difference. As a result of this correlation, codons with \( p_N > p_S \) were characterized not only by unusually high \( p_N \) but also by unusually low \( p_S \). On the other hand, the number of codons with \( p_N > \bar{p}_N \) (where \( \bar{p}_N \) is the mean \( p_N \) for all codons) was very similar to the random expectation, and the observed number of 30-codon windows with \( p_N > p_S \) was significantly lower than the random expectation. These results imply that the occurrence of a certain number of codons or codon windows with \( p_N > p_S \) is expected given the nature of nucleotide substitution and need not imply the action of positive Darwinian selection.

Introduction

A number of statistical methods have been introduced in order to search, without incorporating any a priori knowledge of protein function, for regions of protein-coding genes that are subject to positive Darwinian selection. These methods rely on the theoretical prediction that most nonsynonymous (amino acid altering) nucleotide mutations are deleterious to protein structure and are thus quickly eliminated by purifying selection (Kimura 1977). Thus, in most protein-coding genes, the number of synonymous nucleotide substitutions per synonymous site \( (d_S) \) is expected to exceed the number of nonsynonymous substitutions per nonsynonymous site \( (d_N) \) (Kimura 1977). By contrast, under positive Darwinian selection favoring changes at the amino acid level, \( d_N \) is expected to exceed \( d_S \) (Hughes and Nei 1988). For example, there are widely used methods based on maximum parsimony (Suzuki and Gojobori 1999) or maximum likelihood (Yang et al. 2000) that estimate \( d_S/d_N \) at individual codons and search for those at which \( d_N > d_S \). Similarly, Fares et al. (2002) proposed a method for estimating the \( d_S/d_N \) in sliding windows along coding regions and searching for the optimal window size to detect deviations from the expected pattern. In addition, likelihood ratio methods have been applied to test a model including a category of codons with \( d_N > d_S \) against a model lacking such a category of codons (Yang et al. 2000).

A problem with any such method is that \( d_S \) and \( d_N \) are likely to vary stochastically along the sequence; thus, it is expected that in certain genomic regions, \( d_N \) will exceed \( d_S \) by chance alone, in the absence of positive selection. Available methods do not appear to include any effective controls for such stochastic variation. In order to assess the extent of such stochastic variation in a real data set, we used aligned protein-coding genes from complete genomes of two closely related species of fungi, \( \text{Saccharomyces cerevisiae} \) and \( \text{Saccharomyces paradoxus} \). By comparing the observed distributions of synonymous and nonsynonymous differences in individual codons and in sliding windows with the random expectation, we compared the observed frequencies of codons and windows showing excess numbers of nonsynonymous substitutions with random expectations.

Methods

Sequence data for 4,133 orthologous genes were derived from the aligned genomes of \( \text{S. cerevisiae} \) and \( \text{S. paradoxus} \) (Kellis et al. 2003). We computed the proportion of synonymous nucleotide differences per synonymous site \( (p_S) \) and the number of nonsynonymous differences per nonsynonymous site \( (p_N) \) between the two genomes at individual codons by the Nei and Gojobori (1986) (NG) method. No correction for multiple hits was applied. All such correction formulas are large-sample approximations and thus not applicable in the present case, in which \( p_S \) and \( p_N \) were computed on a codon-by-codon basis. We excluded the codons at which no synonymous sites were defined, i.e., codons at which both sequences included either ATG (Met) or TGG (Trp). We likewise computed sliding windows of 30 codons along each gene in the data set.

Because of the method by which it counts “fractional sites” in the case of twofold degenerate sites, the NG method will tend to underestimate the number of synonymous sites and thus to overestimate \( p_S \) when there is a strong transitional bias (Li 1993). In order to control for this effect, we analyzed separately a data set consisting exclusively of codons at which both genomes used fourfold degenerate codons. In addition, we applied the modified Nei and Gojobori (ModNG) method (Zhang, Rosenberg, and Nei 1998) separately to twofold degenerate codons, to all codons, and to the sliding windows of 30 codons. Whereas the NG method counts the third position of a two-fold degenerate codon as 1/3 synonymous site, the ModNG method counts such a site as \( R/(1 + R) \) synonymous site, where \( R \) is the transition/transversion ratio. We estimated \( R \) from the observed pattern of transitional and transversional differences at fourfold degenerate sites in our data set. Note that, in the case of fourfold degenerate codons, NG and ModNG methods are equivalent.

We compared the actual distribution of \( p_N \), \( p_S \), and their difference with that for 2,000,000 simulated codons.
Each simulated codon was assigned a \( p_N \) value drawn at random (with replacement) from the observed \( p_N \) values and a \( p_S \) value independently drawn at random from the observed \( p_S \) values. A similar procedure was performed to create a data set of 2,000,000 simulated 30-codon windows. In addition, we conducted two computer simulations of codon evolution at 2,000,000 randomly generated codons using the Evolver program (Yang 1997). In each set, protein evolution followed a Dayhoff model, and codons were allowed to evolve 0.4 substitutions per site with a ratio of synonymous to nonsynonymous of 10:1 (consistent to the level of divergence observed in the actual data; table 1). One simulation was run with \( R = 0.5 \) and the other with \( R = 4.5 \) (as observed in the data; see below).

### Results

**Differences at Individual Codons**

The proportions of synonymous differences per synonymous site (\( p_S \)) and the proportion of nonsynonymous differences per nonsynonymous site (\( p_N \)) were computed between *S. cerevisiae* and *S. paradoxus* at 1,993,217 individual codons in 4,133 genes. When \( p_S \) and \( p_N \) were estimated by the NG method, \( p_N \) exceeded \( p_S \) in 125,714 (6.3%) codons; this proportion was not significantly different from that expected on the basis of randomly pairing \( p_S \) and \( p_N \) values (table 1). However, in the present data set, there was evidence that the NG method substantially overestimated \( p_S \). We estimated \( p_S \) codons where both genomes used a twofold degenerate codon (\( N = 782,559 \)) and at codons where both genomes used a fourfold degenerate codon (\( N = 758,893 \)). Mean \( p_S \) was estimated to be over twice as high at twofold degenerate codons as at fourfold degenerate codons (table 1).

We estimated the transition:transversion ratio (\( R \)) from observed differences at fourfold degenerate sites. There were 151,254 transitions and 33,362 transversions, yielding an estimate of \( R = 4.5 \). Using this value in the ModNG method, we obtained a mean \( p_S \) at twofold degenerate codons (0.2178) very close to that seen at fourfold degenerate sites (0.2536) (table 1). When the ModNG method was applied at all codons, \( p_N \) exceeded \( p_S \) in 128,912 codons (fig. 1A and table 1), significantly more frequently than expected by chance (table 1). There were 3,180 more codons with \( p_N > p_S \) than expected from random pairing of \( p_N \) and \( p_N \) values (table 1). The number of excess codons with \( p_N > p_S \) than (3,180) represents 0.16% of all codons compared and 2.4% of all codons with \( p_N > p_S \) by the ModNG method. When twofold degenerate codons were considered separately, the observed number of codons with \( p_N > p_S \) was significantly greater than expected, whether the NG or ModNG method was used (table 1).

### Relationship Between Synonymous and Nonsynonymous Substitution

One factor that might explain the apparent excess of codons with \( p_N > p_S \) would be a negative correlation between \( p_S \) and \( p_N \). In fact, there was a strong negative correlation (\( r = -0.786; P < 0.001 \)) between \( p_S \) and \( p_N \) at codons where at least one nucleotide difference occurred.
The negative correlation between \( p_S \) and \( p_N \) in synonymous differences but no synonymous differences. In many cases, \( p_N \) may exceed \( p_S \) largely because \( p_S \) happens to be low, as a result of stochastic error, rather than because \( p_N \) is unusually elevated. We tested this interpretation by computer simulations, assuming only purifying selection and a rate of substitution comparable to that observed in the actual data; we observed correlations very similar to those seen in the actual data. For simulations with \( R = 0.5 \) and \( R = 4.5, p_S \) and \( p_N \) at codons where at least one nucleotide difference occurred were negatively correlated (\( r = -0.601 \) and \( r = -0.574 \), respectively; \( P < 0.001 \) in both cases).

**Average Numbers of Nucleotide Differences**

In order to provide further evidence regarding the relationship between synonymous and nonsynonymous substitution, we computed mean values of \( p_S \) and \( p_N \) separately for codons with \( p_N > p_S \) and for codons with \( p_S \leq p_N \) (table 2). Using both NG and ModNG methods, mean \( p_S \) at codons with \( p_N > p_S \) was significantly lower than mean \( p_S \) at codons with \( p_N \leq p_S \) (table 2). This was true of all codons, fourfold degenerate sites, and twofold degenerate codons (table 2). As expected, using both NG and ModNG methods, mean \( p_N \) at codons with \( p_N > p_S \) was significantly greater than mean \( p_N \) at codons with \( p_N \leq p_S \) (table 2).

These results support the hypothesis that codons with \( p_N > p_S \) frequently showed both lower than average \( p_S \) and higher than average \( p_N \). Thus, a more conservative method of identifying codons with elevated \( p_N \) (as might result from positive selection) would be to compare \( p_N \) in a codon not with \( p_S \) in the same codon but rather with the genome-wide mean \( p_S \) (\( p_S^{\text{gen}} \)). Therefore, we computed mean values of \( p_S \) and \( p_N \) separately for codons with \( p_N > p_S \) and for codons with \( p_N \leq p_S \) (table 2). Mean \( p_S \) values for codons with \( p_N > p_S \) were not significantly greater than the mean \( p_S \) for codons with \( p_N \leq p_S \) except in the case of fourfold degenerate codons (table 2).

In the case of all sites using the NG method and in the case of twofold degenerate sites using the ModNG method,
mean $p_S$ values for codons with $p_N > p_S$ were significantly lower than the mean $p_S$ for codons with $p_N \leq p_S$ (table 2). In the case of all sites using the ModNG method, the mean $p_S$ value for codons with $p_N > p_S$ was not significantly different from the mean $p_S$ for codons with $p_N \leq p_S$ (table 2). On the other hand, mean $p_N$ for codons with $p_N > p_S$ was in all cases significantly greater than mean $p_N$ for codons with $p_N \leq p_S$ (table 2). These results suggest that comparing $p_N$ in an individual codon with $p_S$ provides a way of identifying codons with elevated $p_N$ but without unusually low $p_S$.

We computed the numbers of codons with $p_N > p_S$ and compared the observed values with the expectation derived from randomly pairing $p_S$ and $p_N$ values (table 3). The observed number of codons with $p_N > p_S$ was significantly different from the expected value only in the case of two-fold degenerate codons with the NG method (table 3). When the ModNG method was used, the observed and expected values were nearly identical in every case, and the observed and expected values were also identical for the NG method applied to all sites (table 3).

### Sliding Window Analyses

In sliding windows analyses, the observed number of windows with $p_N > p_S$ was significantly greater than expected whether the NG or ModNG method was used (table 1). It was interesting to observe that the use of ModNG greatly increased both the observed and expected numbers of windows with $p_N > p_S$ (table 1). This no doubt occurred because of the reduction in estimates of $p_S$ using ModNG (table 1). The distribution of $p_N - p_S$ for the sliding window was much smoother and closer to a bell-shaped curve than in the case of individual codons (fig. 1). Windows with $p_N > p_S$ had significantly lower mean $p_S$ than windows with $p_N \leq p_S$ and significantly higher mean $p_N$ than windows with $p_N \leq p_S$ (table 2). However, windows with $p_N > p_S$ had significantly higher mean values of both $p_S$ and $p_N$ than windows with $p_N \leq p_S$ (table 2). Furthermore, the observed number of windows with $p_N > p_S$ was very close to the expected number (table 3).

### Discussion

Comparison of the proportions of synonymous ($p_S$) and nonsynonymous ($p_N$) difference in orthologous genes of two yeast species revealed different patterns depending on the sites analyzed and the method used. The ModNG method showed advantages over the original NG method in analyzing this data set. There was a strong transitional bias, as indicated by a transition:transversion ratio of 4.5:1 at fourfold degenerate sites. The NG method gave an estimate of mean $p_S$ at twofold degenerate codons greater than twice that for fourfold degenerate codons (table 2). By contrast, the ModNG method yielded estimates of $p_S$ that were, on average, very similar for fourfold degenerate codons, twofold degenerate codons, and sliding windows of 30 codons (table 1).

Using the ModNG method, we found a significant excess of codons with $p_N > p_S$ in comparison to the expectation based on random shuffling of $p_S$ and $p_N$ values. It might be concluded that the excess in codons with $p_N > p_S$ is caused by a certain proportion of codons which are subject to positive selection favoring nonsynonymous changes. Even if this interpretation is correct, the number of codons affected would be very few because the apparent excess of codons with $p_N > p_S$ represented only 0.16% of all codons and only 2.4% of codons with $p_N > p_S$.

However, there are problems with the hypothesis that the observed excess of codons with $p_N > p_S$ is due to positive selection. First, there was a strong negative correlation between the number of synonymous differences and the number of nonsynonymous differences at codons with at least one nucleotide difference. Second, codons with $p_N > p_S$ tended to have lower than average values of $p_S$ as well as higher than average values of $p_N$ (table 2). These problems were avoided if $p_N$ in a given codon was compared not with $p_S$ in the same codon but with the mean value of $p_S$ for all codons compared ($\bar{p}_S$). However, in the latter case, there was no evidence that codons with elevated $p_N$ occurred at a greater rate than the random expectation.

When we examined 30-codon sliding windows, we found a certain number with $p_N > p_S$ and a certain number with $p_N > p_S$ (tables 1 and 3). However, in neither case were the observed numbers greater than the expected numbers (tables 1 and 3).

Our results imply that the presence of a certain number of codons and of codon windows with a greater proportion of nonsynonymous than of synonymous nucleotide differences is expected given the pattern of nucleotide substitution. Thus, the existence of a set of such codons or codon windows cannot in itself be taken as evidence of the presence of positive Darwinian selection. As a consequence, our results cast doubt on the validity of methods—whether based on parsimony or likelihood—that...
consider the presence of a set of codons with $d_N > d_S$ as evidence of positive selection. Of course, widely used methods look at the pattern of substitution across a phylogenetic tree, whereas we compared two genomes. However, the problems we observed are expected to be present in phylogenetic trees as well; for example, numbers of synonymous and nonsynonymous differences are likely to show stochastic variation from codon to codon, and there is likely to be a negative correlation between the numbers of synonymous and nonsynonymous differences.

The observed negative correlation between $p_S$ and $p_N$ at codons with at least one nucleotide difference appeared to result mainly from the underrepresentation of codons with both synonymous and nonsynonymous differences. Such codons are lacking because the occurrence of nucleotide differences varies stochastically across codons and because the overall rate of nucleotide substitution is less than one substitution per site. In addition, purifying selection reduces the overall occurrence of nonsynonymous difference, resulting in many codons with synonymous but not nonsynonymous differences. Computer simulation supported the hypothesis that these factors alone are sufficient to explain the observed negative correlation between $p_S$ and $p_N$ at codons with at least one nucleotide difference. However, factors such as codon usage bias (Sharp and Li 1986), which will lead to purifying selection even at certain synonymous sites, may also contribute to maintaining this correlation.

These problems may account in part for the overly liberal nature of the commonly used likelihood ratio tests (LRT) for positive selection (Suzuki and Nei 2004; Zhang 2004). The LRT compares a model including a set of codons with $d_S/d_S > 1.0$ with a model lacking such a set. If the former model gives a better likelihood score, the hypothesis of positive selection is accepted. Our results suggest that this is not a valid test of the hypothesis of positive selection because the existence of a set of codons with $d_S/d_S > 1.0$ is expected even under neutrality. An alternative approach might be to search for codons at which $d_N$ exceeds mean $d_S$ for the entire gene. However, our results (table 3) suggest that even the latter approach would still not provide a valid test of positive selection because it is expected that there will be certain codons where, by chance alone, $d_N$ exceeds mean $d_S$ for the entire gene. Rather, these methods of testing for positive selection might be improved if one were to allow for a set of codons with $d_S/d_S > 1.0$ even in the model without positive selection. This model could be tested against a model in which the number of codons with $d_S/d_S > 1.0$ is greater than a threshold value expected under random substitution.

**Acknowledgment**

This research was supported by grant GM43940 from the National Institutes of Health to A.L.H.

**Literature Cited**


Naruya Saitou, Associate Editor

Accepted February 21, 2005