Significant Impact of Protein Dispensability on the Instantaneous Rate of Protein Evolution

Jianzhi Zhang and Xionglei He
Department of Ecology and Evolutionary Biology, University of Michigan

The neutral theory of molecular evolution predicts that important proteins evolve more slowly than unimportant ones. High-throughput gene-knockout experiments in model organisms have provided information on the dispensability, and therefore importance, of thousands of proteins in a genome. However, previous studies of the correlation between protein dispensability and evolutionary rate were equivocal, and it has been proposed that the observed correlation is due to the covariation with the level of gene expression or is limited to duplicate genes. We here analyzed the gene dispensability data of the yeast *Saccharomyces cerevisiae* and estimated protein evolutionary rates by comparing *S. cerevisiae* with nine species of varying degrees of divergence from *S. cerevisiae*. The correlation between gene dispensability and evolutionary rate, although low, is highly significant, even when the gene expression level is controlled for or when duplicate genes are excluded. Our results thus support the hypothesis of lower evolution rates for more important proteins, a widely used principle in the daily practice of molecular biology. When the evolutionary rate is estimated from closely related species, the ratio between the mean rate of nonessential proteins to that of essential proteins is 1.4. This ratio declines to 1.1 when the evolutionary rate is estimated from distantly related species, suggesting that the importance of a protein may change in evolution, so the dispensability data obtained from a model organism only predicts a short-term rate of protein evolution. A comparison of the fitness contributions of orthologous genes in yeast and nematode supports this conclusion.

Introduction

In an influential article entitled “On some principles governing molecular evolution,” Kimura and Ohta (1974) proposed that functionally less important proteins evolve faster than more important ones in terms of amino acid substitution. This was deduced from the neutral theory of molecular evolution as well as summarized from empirical evidence. In the neutral theory (Kimura 1983), the substitution rate per site \( k \) is identical to the neutral mutation rate per site \( (v_0) \) and can be expressed as \( k = v_0 = v_T f_0 \), where \( v_T \) is the total rate of mutation and \( f_0 \) is the proportion of mutations that are neutral. Thus, the theory predicts a higher \( k \) for less important proteins because \( f_0 \) is likely to be greater for less important proteins. Although Kimura and Ohta (1974) provided several empirical examples supporting their prediction, it was difficult to objectively and quantitatively measure the importance of a protein. Wilson, Carlson, and White (1977) proposed that the substitution rate \( k = PQ \), where \( P \) is the probability that a substitution is compatible with the function of the protein and \( Q (>0) \) is the probability that an organism can survive and reproduce without the protein. \( Q \), also known as protein dispensability, can be experimentally quantified and used as a reasonable measure of protein importance. Wilson, Carlson, and White (1977) thus predicted that dispensable (nonessential or unimportant) proteins tend to evolve faster than indispensable (essential or important) proteins. Most molecular biologists appear to agree that important proteins evolve more slowly. In fact, they consciously or unconsciously apply this principle in their daily practice when they use sequence conservation as an indication of functional importance.

The availability of large gene-knockout data from functional genomic studies has offered the opportunity to test whether protein dispensability and evolutionary rate are indeed correlated at the genome-wide level. This was first attempted by Hurst and Smith (1999). They measured the rate of protein evolution by the ratio of the nonsynonymous nucleotide substitution rate \( (d_N) \) to the synonymous rate \( (d_S) \) between orthologous genes of the mouse and rat and measured protein dispensability using knockout phenotypes of 175 mouse genes. They found that nonessential genes evolve more rapidly than essential genes. Here, essential genes are those that when knocked out lead to lethal or sterile phenotypes, and nonessential genes are all other genes. However, after they excluded 34 nonessential immunity genes, which are likely under positive selection, nonessential genes no longer evolve faster than essential genes. They thus concluded that there is no difference in evolutionary rate between essential and nonessential proteins. Hirsh and Fraser (2001) analyzed the fitness effect caused by gene deletion in the yeast *Saccharomyces cerevisiae* and estimated the rate of protein evolution by comparing orthologous genes of the yeast and nematode *Caenorhabditis elegans*. They found a significant trend that genes with smaller fitness effects evolve faster. They also argued based on a population genetic model that the protein evolutionary rate is correlated with the fitness effect only when the fitness effect is weak (<0.5), and they believed that Hurst and Smith’s failure was due to their inclusion of genes with strong fitness effects such as essential genes. It is known that lowly expressed genes evolve faster than highly expressed genes in yeast, although the exact cause of this relationship is unclear (Pal, Papp, and Hurst 2001). In a reanalysis of the yeast data, Pal, Papp, and Hurst (2003) found that the correlation between the evolutionary rate and fitness effect is no longer significant when the gene expression level is controlled for, suggesting that the correlation between fitness effect and evolutionary rate observed by Hirsh and Fraser (2001) is due to covariation with gene expression. In a response to Pal, Papp, and Hurst (2003), Hirsh and Fraser (2003) claimed that the correlation between evolutionary rate and fitness effect was significant even after they controlled for gene expression, when a larger data set and an improved method were used. However, they did not...
publish evidence supporting their assertion. Yang, Gu, and Li (2003) also reanalyzed the yeast data. Instead of using the *S. cerevisiae*—*C. elegans* comparison to estimate the evolutionary rate as in Hirsh and Fraser (2001), they used the *S. cerevisiae*—*Candida albicans* comparison because the latter species pair is evolutionarily much closer to each other. Interestingly, Yang, Gu, and Li (2003) found that the correlation between the evolutionary rate and fitness effect is limited to duplicate genes and is nonexistent among singleton genes. They, however, did not control for gene expression in their study. Castillo-Davis and Hartl (2003) compared *C. elegans* genes showing embryonic lethality in RNAi experiments (RNAi) with those without RNAi phenotypes. They found that the former group of genes evolve significantly more slowly than the latter group and that both duplicate and singleton genes exhibit this difference. In this analysis, they estimated evolutionary rates of *C. elegans* genes by comparing them with *Caenorhabditis briggsae* orthologs. But, gene expression was again not controlled for. In addition to these eukaryotic studies, the correlation between protein dispensability and evolutionary rate has been examined in prokaryotes. While the initial finding strongly supported the existence of such a correlation in prokaryotes (Jordan et al. 2002), the correlation was found to be no longer significant after gene expression was controlled for (Rocha and Danchin 2004).

Despite intensive investigations in the past few years, it remains unclear whether protein dispensability and evolutionary rate are correlated, particularly among singletons and after gene expression is controlled for. Due to the availability of a limited number of genome sequences, most previous studies used relatively divergent species for the estimation of protein evolutionary rate. It is possible that such a practice contributed to the inconsistent results observed by different researchers. Because protein dispensability is measured in one species, while evolutionary rate is estimated through the comparison of two species and is therefore an average for the period of evolutionary time separating the two species, use of closely related species would increase the power of detecting the effect of dispensability on evolutionary rate, if such an effect indeed exists and the protein evolutionary rate changes over time. Recently, the genomes of over a dozen yeast species have been sequenced, and these species form a nice gradient in terms of their evolutionary distances from *S. cerevisiae* (Wolfe 2004). By analyzing these data, here we show that protein dispensability does affect evolutionary rate, even after we control for gene expression and exclude duplicate genes. However, the effect declines with evolutionary time, and protein dispensability measured in one species does not predict the evolutionary rate of the protein in distantly related species.

**Materials and Methods**

**Genomic Data Used**

The genome sequences of *S. cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces bayanus*, and *Saccharomyces castellii* were downloaded from ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence. The genome sequences of *Candida glabrata*, *Debaryomyces Hansenii*, and *Yarrowia lipolytica* were downloaded from ftp://ftp.ncbi.nih.gov/genbank/organisms/Fungi. The genome sequences of *Kluyveromyces waltii*, *Ashbya gossypii*, *C. albicans*, and *C. elegans* were obtained from http://www.broad.mit.edu/seq/YeastDuplication, ftp://ftp.ncbi.nih.gov/organisms/Fungi, http://sequence.stanford.edu/group/candida/download.html, and http://www.ensembl.org/Download/, respectively. The *S. cerevisiae* transcriptome data set that included almost all predicted genes was downloaded from http://web.wi.mit.edu/young/express/transcriptome.html. In generating this transcriptome data set, Holstege et al. (1998) used high-density oligonucleotide arrays to determine the abundance of mRNAs extracted from mid-log–phase yeast cells cultured in the YPD medium. The *S. cerevisiae* single gene deletion fitness data set generated by Steinmetz et al. (2002) was downloaded from http://www-deletion.stanford.edu/YDPM/YDPM_index.html. Following Gu et al. (2003), the lowest fitness value across five growth conditions (YPD, YPDGE, YPE, YPG, and YPL) was used as the fitness of a gene-deletion strain. In addition, a list of essential genes (Winzeler et al. 1999) was downloaded from http://www-sequence.stanford.edu/group/yeast_deletion_project/essential_ORFs.txt. Both data sets were generated by Ronald Davis’ group at Stanford University and are almost always used together (e.g., Gu et al. 2003). But there are ~1% of genes for which different results were obtained by Winzeler et al. (1999) and Steinmetz et al. (2002). After excluding these genes, a total of 5,724 genes with fitness values were used in our analysis. It should be noted that in Steinmetz et al. (2002), the fitness of a deletion strain was measured by the ratio of its growth rate to the average growth rate of all strains. Thus, some fitness values are higher than 1, and the corresponding fitness effects are lower than 0. Because fitness is a relative value, such measurement is acceptable. The *C. elegans* RNAi phenotype data set generated by Kamath et al. (2003) was downloaded from http://www.nature.com/nature/journal/v421/n6920/supplinfo/nature01278.html.

**Data Analyses**

To identify orthologs, genome-wide all-against-all BlastP (Altschul et al. 1990) searches (Evalue = 10−10) were carried out between yeast *S. cerevisiae* and one of the nine other yeasts or *C. elegans*. A hit was considered valid if the non–self-hits in the genome with the highest scoring alignment was longer than 80% of the length of the two proteins that matched. Reciprocal best hits were defined as orthologs. Transposable elements and mitochondrial genes were excluded from the analysis. A list of orthologous genes was obtained between *S. cerevisiae* and each of the nine yeasts. *Saccharomyces cerevisiae* genes that appeared in all the nine lists were subsequently derived. These *S. cerevisiae* genes and their orthologs in the 9 yeasts were used for the analysis involving only shared orthologs across the 10 yeasts. A gene was defined as a singleton if it did not have duplicate copies in the genome. Operationally, a singleton has no non–self-hits in a genome-wide all-against-all BlastP searches (Evalue = 0.1). Conservatively, a gene was defined as a duplicate gene if it had at least one non–self-hit in genome-wide all-against-all BlastP searches (Evalue = 10−20).
Homologous proteins were aligned by Clustal (Thompson, Higgins, and Gibson 1994), and the DNA sequences were then aligned according to the protein alignment. The number of nonsynonymous substitutions per nonsynonymous site between two sequences \( d_N \) was estimated by the likelihood method using PAML (Yang 1997). Rank correlations and partial rank correlations were conducted as described in Sokal and Rohlf (1995).

Results
Significant Correlation Between Fitness Effect and Evolutionary Rate

We analyzed a data set of protein dispensability derived from a large-scale gene deletion experiment in \( S.\) cerevisiae (Steinmetz et al. 2002). Gene dispensability is measured by the fitness effect of gene deletion, which is 1 minus the fitness of the yeast strain with a specific gene deleted. Gene expression was estimated by the number of mRNA copies per gene per cell in mid-log–phase yeast cells cultured in the YPD medium and was measured using high-density oligonucleotide arrays (Holstege et al. 1998). We first used the \( S.\) cerevisiae—\( S.\) paradoxus comparison to estimate the evolutionary rate of yeast proteins because among the yeasts whose genomes have been completely sequenced, \( S.\) paradoxus is closest to \( S.\) cerevisiae (Wolfe 2004; fig. 1). Using stringent criteria (see Materials and Methods), we identified orthologous genes between the two yeasts. After removing genes with either no gene expression data or no fitness effect data, we obtained 4,201 orthologous gene pairs for further analysis. We measured the rate of protein evolution by the nonsynonymous substitution rate \( d_N \). The average \( d_N \) between the two yeasts is 0.0407 (table 1), which is substantially lower than the corresponding \( d_N \) in virtually all previous studies of the relationship between protein dispensability and evolutionary rate. For example, the average \( d_N \) was about 0.1 in Castillo-Davis and Hartl (2003), 0.3 in Rocha and Danchin (2004), and 0.4–0.5 in Yang, Gu, and Li (2003). Although not presented, the average \( d_N \) in Hirsh and Fraser (2001) and Pal, Papp, and Hurst (2003) would be considerably larger than 0.0407, as the divergence of the species pairs they used is much greater than the divergence between \( S.\) cerevisiae and \( S.\) paradoxus. The average \( d_N \) in Hurst and Smith’s (1999) study is not known. If we use the genome-wide average \( d_S \) of mouse-rat orthologous genes estimated elsewhere, it would be about 0.02 (Gibbs et al. 2004). Although this number is lower than the average \( d_S \) in our comparison, the small size of the Hurst and Smith sample (141 genes) may have rendered their analysis powerless.

We found that \( d_S \) estimated from the \( S.\) cerevisiae—\( S.\) paradoxus comparison and the fitness effect of gene deletion in \( S.\) cerevisiae are negatively correlated (fig. 2; Spearman’s rank correlation \( r_1 = -0.19, n = 4,201, P = 2 \times 10^{-35} \); table 1). \( d_N \) and the level of gene expression are also negatively correlated (Spearman’s rank correlation \( r_2 = -0.51, n = 4,201, P = 4 \times 10^{-273} \); table 1), as previously found (Pal, Papp, and Hurst 2001). Not unexpectedly, the fitness effect and gene expression level are positively correlated (Spearman’s rank correlation \( r_1 = 0.20, n = 4,201, P = 1 \times 10^{-39} \); table 1), meaning that highly expressed genes tend to have greater fitness effects when deleted.

We then performed a partial correlation analysis and found that \( d_N \) and fitness effect are still negatively correlated even after controlling for gene expression (partial rank correlation \( r_1 = -0.10, n = 4,201, P = 1 \times 10^{-14} \); table 1). While highly significant, the level of this correlation is low, as only \( r_N^2 = 0.02 \) of among-gene variation in evolutionary rate is explainable by the variation in fitness effect. The partial correlation between \( d_N \) and gene expression level is also significant after the control for fitness effect (partial rank correlation \( r_2 = -0.49, n = 4,201, P = 8 \times 10^{-59} \); table 1), and the partial correlation between the fitness effect and gene expression level is significant after the control for \( d_N \) (partial rank correlation \( r_3 = 0.12, n = 4,201, P = 8 \times 10^{-16} \); table 1). It is interesting to note that the difference between \( R_1 \) and \( r_1 \) is greater than that between \( R_2 \) and \( r_2 \) (table 1). However, it is more meaningful to compare the difference between \( R_1^2 \) and \( r_1^2 \) with the difference between \( R_2^2 \) and \( r_2^2 \) because the square of the correlation coefficient measures the proportion of variance of one variable that can be explained by the second variable. We found that \( R_1^2 - r_1^2 \) is similar to \( R_2^2 - r_2^2 \).

To examine whether the correlation between \( d_N \) and fitness effect is different for singleton and duplicate genes, we identified singleton and duplicate genes of \( S.\) cerevisiae by all-against-all BlastP searches (see Materials and Methods). Conservatively, singletons are defined as genes that do not have duplicate copies in the genome of \( S.\) cerevisiae when \( E\) value = 0.1 was used as the cutoff in the BlastP searches, whereas duplicate genes are those with at least one non–self-hit in \( S.\) cerevisiae when \( E\) value = \( 10^{-20} \) was used. After removing genes without \( S.\) paradoxus orthologs, gene expression information, or fitness effect data, a total of 1,124 \( S.\) cerevisiae singleton genes and 1,513 duplicate genes were obtained. We found that for both singleton and duplicate genes, the correlation between evolutionary rate (\( d_N \)) and fitness effect is significant (table 2). This correlation remains significant even when we control for gene expression (table 2).

While the correlation coefficient between two variables can be used to measure the influence of one variable on the other, we can also classify genes into groups according to one variable and then study the second variable among groups. This was the strategy used by Hurst and...
Table 1
Correlations and Partial Correlations Among Protein Dispensability, Evolutionary Rate, and Expression Level

<table>
<thead>
<tr>
<th>Species Compared with Saccharomyces cerevisiae</th>
<th>Mean $d_{ab}$ (standard deviation of mean)</th>
<th>$\lambda_1^a$</th>
<th>$\lambda_2^b$</th>
<th>$R_1^c$</th>
<th>$R_2^d$</th>
<th>$R_3^e$</th>
<th>$r_1^f$</th>
<th>$r_2^g$</th>
<th>$r_3^h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saccharomyces paradoxus</td>
<td>4.201 (0.001)</td>
<td>1.39 (1 x 10^-30)</td>
<td>1.46 (7 x 10^-36)</td>
<td>-0.19 (2 x 10^-35)</td>
<td>-0.51 (4 x 10^-23)</td>
<td>0.20 (1 x 10^-30)</td>
<td>-0.10 (1 x 10^-11)</td>
<td>-0.49 (1 x 10^-24)</td>
<td>0.12 (8 x 10^-15)</td>
</tr>
<tr>
<td>Saccharomyces bayanus</td>
<td>3.751 (0.008)</td>
<td>1.43 (1 x 10^-40)</td>
<td>1.52 (2 x 10^-36)</td>
<td>-0.23 (2 x 10^-38)</td>
<td>-0.42 (4 x 10^-20)</td>
<td>0.21 (2 x 10^-38)</td>
<td>-0.15 (2 x 10^-19)</td>
<td>-0.50 (5 x 10^-23)</td>
<td>0.11 (5 x 10^-16)</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>3.011 (0.003)</td>
<td>1.32 (7 x 10^-27)</td>
<td>1.37 (7 x 10^-32)</td>
<td>-0.22 (6 x 10^-33)</td>
<td>-0.46 (3 x 10^-24)</td>
<td>0.18 (2 x 10^-34)</td>
<td>-0.15 (2 x 10^-21)</td>
<td>-0.54 (7 x 10^-22)</td>
<td>0.07 (1 x 10^-4)</td>
</tr>
<tr>
<td>Candida glabrata</td>
<td>3.249 (0.004)</td>
<td>1.25 (2 x 10^-18)</td>
<td>1.31 (4 x 10^-22)</td>
<td>-0.20 (5 x 10^-32)</td>
<td>-0.60 (1 x 10^-26)</td>
<td>0.18 (3 x 10^-25)</td>
<td>-0.12 (2 x 10^-12)</td>
<td>-0.58 (3 x 10^-29)</td>
<td>0.07 (2 x 10^-5)</td>
</tr>
<tr>
<td>Kluyveromyces waltii</td>
<td>2.890 (0.004)</td>
<td>1.25 (7 x 10^-21)</td>
<td>1.30 (7 x 10^-26)</td>
<td>-0.21 (8 x 10^-31)</td>
<td>-0.57 (2 x 10^-24)</td>
<td>0.17 (7 x 10^-20)</td>
<td>-0.14 (8 x 10^-15)</td>
<td>-0.55 (1 x 10^-23)</td>
<td>0.06 (1 x 10^-4)</td>
</tr>
<tr>
<td>Ashbya gossypii</td>
<td>2.785 (0.002)</td>
<td>1.27 (1 x 10^-36)</td>
<td>1.34 (8 x 10^-36)</td>
<td>-0.25 (8 x 10^-41)</td>
<td>-0.59 (3 x 10^-25)</td>
<td>0.19 (7 x 10^-25)</td>
<td>-0.18 (2 x 10^-21)</td>
<td>-0.57 (7 x 10^-28)</td>
<td>0.05 (1 x 10^-4)</td>
</tr>
<tr>
<td>Debaryomyces hansenii</td>
<td>2.110 (0.005)</td>
<td>1.19 (8 x 10^-18)</td>
<td>1.25 (1 x 10^-20)</td>
<td>-0.22 (9 x 10^-28)</td>
<td>-0.84 (9 x 10^-15)</td>
<td>0.21 (3 x 10^-23)</td>
<td>-0.13 (7 x 10^-9)</td>
<td>-0.54 (3 x 10^-15)</td>
<td>0.11 (3 x 10^-3)</td>
</tr>
<tr>
<td>Candida albicans</td>
<td>1.892 (0.008)</td>
<td>1.05 (1 x 10^-10)</td>
<td>1.18 (2 x 10^-13)</td>
<td>-0.20 (1 x 10^-17)</td>
<td>-0.08 (4 x 10^-5)</td>
<td>0.17 (1 x 10^-4)</td>
<td>-0.06 (1 x 10^-3)</td>
<td>-0.50 (9 x 10^-14)</td>
<td>0.12 (1 x 10^-4)</td>
</tr>
<tr>
<td>Yarrowia lipolytica</td>
<td>1.610 (0.006)</td>
<td>1.11 (3 x 10^-5)</td>
<td>1.15 (5 x 10^-8)</td>
<td>-0.17 (9 x 10^-12)</td>
<td>-0.52 (2 x 10^-11)</td>
<td>0.23 (3 x 10^-20)</td>
<td>-0.06 (2 x 10^-2)</td>
<td>-0.50 (2 x 10^-10)</td>
<td>0.17 (2 x 10^-11)</td>
</tr>
</tbody>
</table>

$^a$ The mean evolutionary rate ($d_{ab}$) of genes with nonlethal phenotypes divided by that of genes with lethal phenotypes when deleted in S. cerevisiae. In the parentheses is the probability of the observation under the hypothesis of $\lambda_1 = 1$.

$^b$ The mean evolutionary rate ($d_{ab}$) of genes with weak fitness effects ($<5\%$) divided by that of genes with lethal phenotypes when deleted in S. cerevisiae. In the parentheses is the probability of the observation under the hypothesis of $\lambda_2 = 1$.

$^c$ Spearman’s rank correlation between fitness effect and $d_{ab}$. In the parentheses is the probability of the observation under the hypothesis of no correlation.

$^d$ Spearman’s rank correlation between expression level and $d_{ab}$. In the parentheses is the probability of the observation under the hypothesis of no correlation.

$^e$ Spearman’s rank correlation between fitness effect and expression level. In the parentheses is the probability of the observation under the hypothesis of no correlation.

$^f$ Partial rank correlation between fitness effect and $d_{ab}$ when expression level is controlled for. In the parentheses is the probability of the observation under the hypothesis of no correlation.

$^g$ Partial rank correlation between expression level and $d_{ab}$ when fitness effect is controlled for. In the parentheses is the probability of the observation under the hypothesis of no correlation.

$^h$ Partial rank correlation between fitness effect and expression level when both effects are controlled for. In the parentheses is the probability of the observation under the hypothesis of no correlation.
rate, which is actually the average rate during the divergence of the two closely related species. We repeated the rate estimation using comparisons between *S. cerevisiae* and each of eight more divergent species of yeasts (fig. 1) and studied the influence of fitness effect on evolutionary rate. We found that for each of these eight comparisons, protein dispensability as measured by fitness effect of gene deletion has a small, yet statistically significant, impact on the rate of protein evolution (dN), even after controlling for gene expression (table 1). Both the partial correlation between dN and gene expression after the control for fitness effect and the partial correlation between the fitness effect and gene expression after the control for dN remain significant for each of the eight species considered (table 1).

To investigate how the level of species divergence affects the degree to which protein dispensability impacts the average rate of protein evolution, we plotted λ₁ and λ₂ against the mean dN between species pairs for which the average evolutionary rates were estimated. The mean dN was computed by considering all orthologous genes (singleton and duplicate genes) between a species pair. Figure 3 shows that λ₁ and λ₂ are both higher than 1 for all species considered. More importantly, there is a clear trend that both λ₁ and λ₂ decline as the mean dN between species increases, indicating that the impact of protein dispensability on evolutionary rate reduces with evolutionary time. While the dispensability data from the *S. cerevisiae* might predict the average evolutionary rate between *S. cerevisiae* and *S. paradoxus* quite well, it does not predict the average rate between *S. cerevisiae* and *Y. lipolytica* so well. This is likely due to changes in protein function, dispensability, and evolutionary rate over a long evolutionary time, even for orthologous genes.

When two sequences are highly divergent, accurate estimation of dN becomes difficult due to multiple substitutions, which may reduce the observed correlation between dN and protein dispensability. In our analysis, however, the highest mean dN in figure 3 is about 0.5, and the likelihood estimation of dN should be reliable at this level. One caveat in the above analysis is that different genes were used in the comparisons of different species pairs. Because different genes may have different levels of rate constancy over time, a more direct analysis would be to use the same set of genes for all the comparisons. We obtained 680 *S. cerevisiae* genes that have orthologs in each of the other nine yeasts. Figure 4 shows that for these genes, λ₁ and λ₂ decline as species divergence increases, similar to what we observed in figure 3. Consistently, for these 680 genes, the level of correlation between evolutionary rate and fitness effect declines as the species compared diverge from *S. paradoxus* (R₁ = -0.16, P = 3 × 10⁻⁷) to *Y. lipolytica* (R₁ = -0.04, P = 0.002).

One may still argue that the observation of lower λ₁ and λ₂ for more distantly related species could be caused by the confounding effect of gene duplication. For example, it is possible that a *S. cerevisiae* gene has multiple *Y. lipolytica* orthologs due to gene duplication in *Y. lipolytica* after the divergence of the two species. Although we measure the evolutionary rate between the *S. cerevisiae* gene with one of its *Y. lipolytica* orthologs, the fitness effect measured in *S. cerevisiae* would be a poor

| Table 2: Correlations and Partial Correlations Among Protein Dispensability, Evolutionary Rate, and Expression Level for Duplicate and Singleton Genes |
|-----------------|----------------|----------------|----------------|
| Genes | Number of Orthologs | dN (mean) | dN (deviation of mean) |  λ₁ |  λ₂ |  R₁ |  R₂ |  R₃ |  R₄ |
| Singleton | 1,144 | 1.24 | 0.034 (0.001) | 1.26 | 1.01 | 0.94 | 0.96 | 0.96 | 0.96 |
| Duplicate | 1,144 | 1.24 | 0.034 (0.001) | 1.26 | 1.01 | 0.94 | 0.96 | 0.96 | 0.96 |

Note: The evolutionary rate is computed by using the orthologs of *S. cerevisiae* and *S. paradoxus*. The mean evolutionary rate (dN) of genes with nonlethal phenotypes when deleted in *S. cerevisiae* is 1.26 (P = 3 × 10⁻⁷). In the parentheses is the probability of the observation under the hypothesis of no correlation.
Phenotypic Effects of Gene Deletion in Yeast and RNAi in Nematode

To examine whether the function and importance of a gene may change during evolution, we compared the phenotypic effects of gene deletion in the yeast and RNAi in the nematode. The genome-wide RNAi phenotype data were generated by Kamath et al. (2003). In that work, the authors fed nematode *C. elegans* with bacteria expressing double-stranded RNA that correspond to nematode functional genes. Because protein production is inhibited by RNA interference in a gene-specific manner, RNAi mimics the effect of gene deletion. However, RNAi is not always effective, meaning that sometimes protein production may not be effectively inhibited. Using stringent criteria, we identified 735 pairs of orthologous genes between *S. cerevisiae* and *C. elegans*. Interestingly, among 472 genes with no RNAi phenotypes in the nematode, 139 (29%) have lethal phenotypes in yeast. Of 287 genes with lethal effect in yeast, 139 (48%) have no RNAi phenotypes in nematode. Although these cases may in part be due to a low efficiency of RNAi, the following comparison should be biologically meaningful. That is, of 259 genes with less than 5% fitness effect in yeast, 16 (6%) cause 100% embryonic lethality or sterility in nematode. On the other hand, of 137 genes causing 100% embryonic lethality or sterility in nematode, 17 (12%) have less than 5% fitness effect in the yeast. Although the above two percentages are not high, we note that these reflect extreme cases of alteration of dispensability between yeast and nematode orthologs, and there are probably many more genes with mild changes in gene function and dispensability.

**Discussion**

In this work, we took advantage of the recently determined genome sequences of multiple yeast species and studied how protein dispensability affects the rate of protein evolution. Our results support the hypothesis that important proteins evolve more slowly than less important proteins, even when gene expression is controlled for. Although
we could detect statistically significant correlation when the evolutionary rate is estimated by comparing *S. cerevisiae* with any of the nine yeasts considered, the impact of protein dispensability on evolutionary rate declines as the species pair considered becomes more divergent. For instance, the average evolutionary rate of proteins with nonlethal effects is \( \sim 1.4 \) times that of proteins with lethal effects when we estimate the average evolutionary rate between the closely related species of *S. cerevisiae* and *S. paradoxus* (mean \( d_N = 0.04 \)). The corresponding number reduces to \( \sim 1.1 \) when the rate was estimated from the distantly related *S. cerevisiae* and *Y. lipolytica* (mean \( d_N = 0.51 \)). Figure 3 shows a significant linear correlation between \( \lambda_1 \) and mean \( d_N \) (\( R = 0.942, P = 0.0001 \)). Based on this linear regression, it is predicted that \( \lambda_1 \) approaches 1.46 when \( d_N \) approaches 0. In other words, the instantaneous evolutionary rate of *S. cerevisiae* proteins with nonlethal effects is 1.46 times that of proteins with lethal effects. Similarly, figure 3 shows a significant linear correlation between \( \lambda_2 \) and mean \( d_N \) (\( R = 0.934, P = 0.0002 \)), and it may be predicted based on the linear regression that \( \lambda_2 \) approaches \( 1.55 \) when \( d_N \) approaches 0. That is, the instantaneous evolutionary rate of *S. cerevisiae* proteins with weak fitness effects is \( 1.55 \) times that of proteins with lethal effects. These results demonstrate that protein dispensability has a significant impact on the instantaneous rate of protein evolution. However, this impact reduces with evolutionary time, and protein dispensability is a poor predictor of the long-term rate of protein evolution. This is probably because protein importance and evolutionary rate change through time, even for orthologous genes. Our analysis of the phenotypes of orthologous gene deletion in the yeast and RNAi in the nematode is consistent with this view.

Our findings imply that protein dispensability measured in *S. cerevisiae* does not predict the rate of protein evolution in nematodes or other species that are distantly related to the yeast, contradictory to what Hirsh and Fraser (2001) claimed. Their results were based on a small set of genes (119), and it is possible that the correlation they observed was accidental, as suggested by Pal, Papp, and Hurst (2003). Furthermore, in contrast to what Hirsh and Fraser (2001) hypothesized, we found that the correlation between protein dispensability and evolutionary rate can be demonstrated without removing genes of great fitness effects. For instance, we found that the average evolutionary rate for proteins of nonlethal effects is \( 40\% \) greater than that of proteins of lethal effects when closely related species are compared. Following Hirsh and Fraser (2001), we also analyzed a subset of genes with fitness effects lower than 0.5 but found that the correlation between protein dispensability and evolutionary rate is lower for this subset than for the entire data set. For example, when the *S. cerevisiae*–*S. paradoxus* comparison was used for estimating the evolutionary rate, the rank correlation between fitness effect and evolutionary rate was 0.19 (\( P = 2 \times 10^{-15} \)) for the entire data set but only 0.10 (\( P = 5 \times 10^{-9} \)) for the subset of genes with low fitness effects. When we controlled for gene expression, the partial rank correlation between fitness effect and evolutionary rate decreased from 0.10 (\( P = 1 \times 10^{-11} \)) for the entire data set to 0.06 (\( P = 6 \times 10^{-4} \)) for the subset. This partial rank correlation is no longer statistically significant for the subset (\( r = 0.02, P = 0.42 \)) when divergent species such as *C. albicans* is compared with *S. cerevisiae*, although it remains significant for the entire data set (\( r = 0.08, P = 4 \times 10^{-3} \)). Thus, opposite to what Hirsh and Fraser (2001) proposed, our results showed that the effect of protein dispensability on evolutionary rate is less obvious when only genes of low fitness effects are considered. Use of the subset of genes instead of the entire data set was likely the reason why Pal, Papp, and Hurst (2003) could not detect significant impact of protein dispensability on evolutionary rate when gene expression was controlled for. From these considerations, we believe that the findings of Hirsh and Fraser (2001) were by chance, and the evolutionary model they proposed to explain the observation was either unrealistic or irrelevant. Their explanation of why Hurst and Smith (1999) failed to detect the correlation between protein dispensability and evolutionary rate in rodents is probably incorrect as well. We believe that the correlation will be found for rodents when a larger data set is used, unless what we demonstrated in yeasts does not apply to mammals, which seems unlikely.

We detected significant impact of protein dispensability on evolutionary rate for both duplicate and singleton genes. The impact is greater for duplicates than for singletons, as observed by Yang, Gu, and Li (2003) in yeasts and Castillo-Davis and Hartl (2003) in nematodes. The cause of this phenomenon is unclear. Yang, Gu, and Li (2003) suggested that duplicates are to some extent redundant in function, and both the fitness effect and evolutionary rate of a duplicate gene are affected by the level of functional
redundancy that the gene shares with its duplicate copy, generating a correlation between the fitness effect and evolutionary rate. However, functional redundancy can also occur between nonparalogous genes. Furthermore, duplicate genes change functions and rates more rapidly than singletons. Thus, it is puzzling why the impact of protein dispensability on evolutionary rate is higher for duplicates than for singletons.

As found by Pal, Papp, and Hurst (2001), our analysis showed that highly expressed genes have low rates of evolution. This correlation is much stronger than the correlation between fitness effect and evolutionary rate, although the former correlation cannot fully explain the latter. The phenomenon of low evolutionary rates for highly expressed genes has also been documented in bacteria (Rocha and Danchin 2004), plants (Wright et al. 2004), and animals (e.g., Duret and Mouchiroud 2000; Subramanian and Kumar 2004), but the underlying cause remains unclear. As is shown in table 1, functional importance only explains a small fraction of the correlation between expression level and evolutionary rate. If different amino acids are synthesized with different costs or incorporated into a peptide with different rates and accuracies during translation, one may hypothesize that certain amino acids would be preferentially used in highly expressed genes (Akashi and Gojobori 2002; Akashi 2003). This would generate an amino acid usage bias in a way similar to the frequently observed codon usage bias. As the codon usage bias leads to the reduction of the synonymous substitution rate (Sharp and Li 1987), the amino acid bias can reduce the rate of amino acid substitution. Consistent with this hypothesis, biased usage of amino acids has been reported in highly expressed genes (Akashi and Gojobori 2002; Akashi 2003; Urrutia and Hurst 2003; Comeron 2004; Rocha and Danchin 2004).

However, the level of this bias does not seem to fully explain the high correlation of expression level and evolutionary rate (Rocha and Danchin 2004). Another hypothesis is that highly expressed genes may have low mutation rates because of transcription-coupled repair (Svejstrup 2002). This would reduce substitution rates at synonymous, non-synonymous, and intron sites. However, Duret and Mouchiroud (2000) found no reduction in mutation rate in genes expressed in the germ line, contradictory to the prediction of the above hypothesis. It is likely that the correlation of protein evolutionary rate and expression level has multiple causes, but the major cause has yet to be identified. Another interesting question is whether the impact of expression level on evolutionary rate is transient, as observed for the impact of protein dispensability on evolutionary rate. Given the rapid evolution of gene expression patterns (Khaitovich et al. 2004; Yanai, Graur, and Ophir 2004), this prediction seems reasonable. We are currently testing this and other hypotheses in an attempt to understand the strong impact of gene expression on the rate of protein evolution.

Acknowledgments

We thank Wendy Grus and Ondrej Podlaha for valuable comments. This work was supported in part by the NIH research grant GM67030 to J.Z.

Literature Cited


Dan Graur, Associate Editor

Accepted January 25, 2005