Rate of Protein Evolution Versus Fitness Effect of Gene Deletion

Jing Yang, Zhenglong Gu, and Wen-Hsiung Li

Department of Ecology and Evolution, The University of Chicago

Whether nonessential genes evolve faster than essential genes has been a controversial issue. To resolve this issue, we use the data from a nearly complete set of single-gene deletions in the yeast *Saccharomyces cerevisiae* to assess protein dispensability. Also, instead of the nematode, which was used previously but is only distantly related to *S. cerevisiae*, we use another yeast, *Candida albicans*, as a second species to estimate the evolutionary distances between orthologous genes in two species. Our analysis reveals only a weak correlation between protein dispensability and evolutionary rate. More important, the correlation disappears when duplicate genes are removed from the analysis. And surprisingly, the average rate of nonsynonymous substitution is considerably lower than that for single-copy genes in the yeast genome. This observation suggests that structural constraints are more important in determining the rate of evolution of a protein than dispensability because duplicate genes are on average more dispensable than single-copy genes. For duplicate genes, those with only a weak effect or no effect of deletion on fitness evolve on average faster than those with a moderate or strong effect of deletion on fitness, which in turn evolve on average faster than those with a lethal effect of deletion.

Introduction

Wilson, Carlson, and White (1977) predicted that proteins that differ in dispensability are subject to different levels of purifying selection and will evolve at different rates. By analyzing 108 nonessential genes and 67 essential genes in mice that were inferred from knockout studies, Hurst and Smith (1999) found that the rate of nonsynonymous substitution in a gene is not correlated with the severity of the knockout phenotype. Here a gene is considered to be essential if a knockout results in (conditional) lethality or infertility, but nonessential if the knockout yields a viable and fertile individual. Hurst and Smith (1999) argued that although at first sight nonessential genes appeared to be evolving faster than essential ones, the nonessential gene class contains a high proportion of immune-system genes, which tend to evolve fast because they are probably under directional selection caused by host-parasite coevolution. However, Hirsh and Fraser (2001) found a significant negative correlation between evolutionary rate and fitness effect of gene deletion when the fitness effect (reduction) was restricted to the range from 0 to ~0.4. They used the data from a parallel growth assay of single-gene deletions in the yeast to assess protein dispensability (Winzeler et al. 1999) and used *Caenorhabditis elegans* as a second species to compute evolutionary distances and relative evolutionary rates. Moreover, in a more recent study, Jordan et al. (2002) found that essential bacterial genes have been better conserved than nonessential genes in three bacterial species: *Escherichia coli*, *Helicobacter pylori*, and *Neisseria meningitidis*. As the conclusions from these latter two studies conflict with that of Hurst and Smith (1999), this issue deserves further studies. We note that to estimate the evolutionary rates for genes, Hirsh and Fraser (2001) used the nematode as a second species, which is very distantly related to *Saccharomyces cerevisiae*. To obtain more reliable evolutionary distance estimates, we shall use another yeast species, *Candida albicans*. We also note that some of the genes used by Jordan et al. (2002) did not have knockout data and their degrees of dispensability were indirectly inferred from existing data on functional characterizations. In our study, we shall take advantage of an extensive data set of the growth effect of gene deletion in *S. cerevisiae*, which has recently become available (Steinmetz et al. 2002) and can be used as a reliable source for classifying genes into different fitness effect groups.

Key words: protein dispensability, functional compensation, nonsynonymous rate, duplicate genes, singletons.

E-mail: whli@uchicago.edu.

Mol. Biol. Evol. 20(5):772–774. 2003
DOI: 10.1093/molbev/msg078

Materials and Methods

Saccharomyces cerevisiae sequence data were from SGD (Saccharomyces Genome Database, http://genome-www.stanford.edu/Saccharomyces/). We use the fitness data from a nearly complete set of single gene deletions in *S. cerevisiae* (Winzeler et al. 1999; Steinmetz et al. 2002) to estimate the effect of single-gene deletion. The fitness value f_i is defined as r_i / r_{pool}, where r_i is the growth rate of the strain with gene i deleted and r_{pool} is the pooled average growth rate of different strains (Steinmetz et al. 2002). Five growth media were studied: YPD (1% Bactopeptone [Difco], 2% yeast extract and 2% glucose), YPDGE (0.1% glucose, 3% glycerol, and 2% ethanol), YPE (2% ethanol), YPG (3% glycerol), and YPL (2% lactate). We classify yeast genes into three groups according to their fitness values (f) on all studied conditions: (1) If $f > 0.95$ for all five media conditions, the deletion has a weak or no fitness effect in all conditions. (2) If $0 < f_{min} < 0.95$, where f_{min} is the smallest f value for all five growth conditions, the deletion has a moderate to strong effect. (3) If the deletion is lethal, we set $f = 0$.

We use another yeast, *C. albicans*, instead of the nematode used by Hirsh and Fraser (2001), as a second species, so that more accurate estimates of evolutionary distances between species can be obtained. The *Candida* sequences were from the Stanford DNA Sequencing and Technology Center (http://www.sequence.stanford.edu/group/candida/). The orthologous gene pairs between these two yeast species were identified by reciprocal best hits (RBHs) with the total length of alignable regions >80% of the longer protein and $E < 10^{-10}$ in reciprocal search using FASTA (Pearson and Lipman 1988).

Table 1
Comparison of Average K_A's for Different Fitness Groups

Gene Group	Fitness Value	No. of Genes	Average K_A[a]	SD[b] of K_A
Lethal	0	532	0.42	0.22
Strong or moderate effect	0–0.95	452	0.46	0.22
Weak or no effect	$\geqslant 0.95$	880	0.50	0.20
After exclusion of ribosomal proteins				
Lethal	0	520	0.43	0.22
Strong or moderate effect	0–0.95	414	0.47	0.21
Weak or no effect	$\geqslant 0.95$	875	0.50	0.20

[a] A Kruskal-Wallis test reveals a significant rate heterogeneity among the three gene groups $P < 10^{-13}$. A Wilcoxon–Mann-Whitney U test is used to test the difference of average K_A between two groups. The genes with a weak or no effect of deletion on fitness have a significantly higher average K_A than those with a strong or moderate effect of deletion on fitness ($P < 0.001$), which in turn have a significantly higher average K_A than those with a lethal effect of deletion ($P = 0.001$). After the exclusion of ribosomal proteins, the same conclusions hold at the $P < 0.05$ level.

[b] SD: standard deviation.

From 5,919 yeast open reading frames (ORFs) for which we have fitness data, we find 1,864 ORFs that have orthologous genes in *C. albicans*. Protein alignments are conducted using ClustalW (Thompson, Higgins, and Gibson 1994). We then use the PAML package with default parameters (Yang and Nielsen 2000) to calculate the number of substitutions per nonsynonymous site (K_A) and the number of substitutions per synonymous site (K_S) as measures of relative substitution rates.

We also compared the rates of evolution in single-copy genes (singletons) and duplicate genes. A singleton is defined as a protein that did not hit any other proteins in the yeast genome in the FASTA search with E = 0.1; this loose similarity search criterion was used to make sure that a singleton is indeed a singleton. Duplicate genes were identified as in Gu et al. (2002) except that the criterion of 80% alignable regions between protein sequences was reduced to 50% alignable regions because the 80% requirement tends to miss some duplicate genes.

Because ribosomal proteins are known to evolve slowly and the yeast genome contains many ribosomal protein genes, we present our analysis with and without the ribosomal proteins to see if these genes affect our conclusion (tables 1 and 2).

Results and Discussion

Table 1 shows that there is a weak tendency for the average K_A to increase with the fitness value of the deletion strain. Indeed, the group of genes with a weak or no fitness effect of deletion has a significantly higher average K_A value than the group with a strong or moderate fitness effect of deletion, which in turn has a significantly higher average K_A than the lethal group. This conclusion holds regardless of whether the ribosomal proteins are included in the analysis. This result suggests that there is a weak relationship between rate of evolution and protein dispensability, which is in agreement with the results obtained by Hirsh and Fraser (2001) and Jordan et al. (2002). The K_A difference between the two extreme groups, the lethal group and the weak or no effect group, is about 20%, which can be taken as the upper limit of the average effect of dispensability on K_A.

Table 2
Comparisons of Average K_A's for Different Fitness Groups Between Singletons and Duplicate Genes

	Singletons			Duplicate Genes		
Fitness Value	No. of Genes	Ave. K_A[a]	SD[b] of K_A	No. of Genes	Ave. K_A[c]	SD[b] of K_A
0	223	0.50	0.22	104	0.26	0.15
0–0.95	171	0.51	0.21	117	0.32	0.18
$\geqslant 0.95$	306	0.53	0.19	271	0.41	0.18
After exclusion of ribosomal proteins						
0	217	0.50	0.22	102	0.26	0.14
0–0.95	159	0.51	0.22	98	0.36	0.17
$\geqslant 0.95$	306	0.53	0.19	268	0.42	0.18

NOTE.—The total number of genes is now 1,192, compared to 1,864 in table 1. This reduction is due to the definitions of singletons and duplicated genes we described in *Materials and Methods*. As our classification of genes into singletons and duplicated genes is very stringent, we exclude many ambiguous cases.

[a] A Kruskal-Wallis test reveals no heterogeneity of average K_A among the three different fitness groups of singletons: the P values are 0.28 and 0.32 before and after the exclusion of ribosomal proteins.

[b] SD = standard deviation.

[c] A Kruskal-Wallis test reveals significant heterogeneity of average K_A among the three different fitness groups of duplicate genes: the P values are $<10^{-15}$ and $<10^{-15}$ before and after the exclusion of ribosomal proteins. The Wilcoxon–Mann-Whitney U test is used to test the difference in average K_A between two groups. The duplicate genes with a weak or no effect of deletion on fitness have a significant higher average K_A than those with a strong or moderate effect of deletion on fitness ($P < 10^{-6}$), which in turn have a significantly higher average K_A than those with a lethal effect of deletion ($P = 0.008$). The same conclusion holds when ribosomal proteins are excluded from the analysis: the two P values become $P = 0.003$ and $P < 10^{-6}$, respectively.

When we divide the genes under study into duplicate genes and singletons (Gu et al. 2002), two different patterns of the relationship between fitness effect of gene deletion and rate of evolution are found (table 2). For singletons, there is no significant relationship between rate of evolution and protein dispensability (the Kruskal-Wallis test, $P > 0.05$), either before or after the ribosomal protein genes are excluded from comparison. In contrast, for duplicate genes there is a clear tendency for the average K_A to increase with the fitness value, regardless of whether the ribosomal protein genes are included (table 2). In particular, the K_A value for the group with a weak or no fitness effect of deletion is 1.5 times higher than that for the lethal group. Therefore, the rate increase with fitness value in table 1 results in large part from the rate increase in the corresponding duplicate gene groups.

Another surprising finding is that the duplicate genes in table 2 have, on average, evolved more slowly than the singletons; that is, they have a lower average K_A value than the singletons. This observation may be explained by assuming that the duplicate genes included in table 2 have, on average, been subject to more stringent selective constraints than the singletons and have thus evolved more slowly. Thus, the observation suggests that structural constraints on proteins may be more important than dispensability because duplicate genes are more dispensable than single-copy genes (Gu et al. 2003). An additional assumption is that the majority of these duplicate genes had passed the stage of functional relaxation, so that they have been subject to purifying selection for some time and thus have not evolved at a fast rate. For this it is interesting to note the finding by Lynch and Conery (2000) that, although most gene duplicates experience a phase of relaxed selection or even accelerated evolution at nonsynonymous sites during the early stage of divergence, they are nevertheless subject to purifying selection in later stages of divergence. Because over 95% of the duplicate gene pairs under study have a K_S value larger than 1.0, most of them are old. Therefore, most of these duplicated genes appear to have passed the early evolutionary stage following gene duplication.

We now consider why duplicate genes show an increase in rate of evolution with dispensability—i.e., with a decreased fitness effect of deletion (table 2). For those duplicated genes with a lethal effect of deletion, the functional divergence is so large that a deleterious mutation in one gene cannot be compensated by the other copy. In contrast, for those duplicate genes with a moderate or strong fitness effect, the functional divergence is on average smaller, so that there is a chance for compensation for mutations. For this reason, their nonsynonymous rate (K_A) has been accelerated to some extent compared to the rate for the lethal group. For duplicate genes with a weak or no fitness effect, the chance for mutation compensation is even higher and so is the degree of acceleration in K_A.

An intriguing question remains: Why have not "dispensable" (or nonessential) singletons evolved faster than nondispensable singletons? This may be because these dispensable singletons are not really dispensable in long-term evolution. As noted by Brookfield (1992), a gene may be dispensable in an individual, but may not be truly dispensable in evolution. In the case of duplicate genes, however, a duplicate copy can be truly or nearly dispensable in evolution if the other copy alone is sufficient to maintain the function.

In conclusion, for a duplicate gene, the rate of evolution may increases with its dispensability—i.e., the fitness value of the deletion strain. In contrast, for single-copy genes there is no clear relationship between dispensability (fitness effect of deletion) and rate of evolution. This conclusion is different from that of Hirsh and Fraser (2001) and Jordan et al. (2002).

Acknowledgments

This study was supported by National Institutes of Health grant GM30998. We thank Lars M. Steinmetz for his help.

Literature Cited

Brookfield, J. 1992. Can genes be truly redundant? Curr. Biol. **2**:553–554.

Gu, Z., A. Cavalcanti, F.-C. Chen, P. Bouman, and W.-H. Li. 2002. Extent of gene duplication in the genome of Drosophila, nematode, and yeast. Mol. Biol. Evol. **19**:256–262.

Gu, Z., L. M. Steinmetz, X. Gu, C. Scharfe, R. W. Davis, and W.-H. Li. 2003. Role of duplicate genes in genetic robustness against null mutations. Nature. **421**:63–66.

Hirsh, A. E., and H. B. Fraser. 2001. Protein dispensability and rate of evolution. Nature **441**:1046–1049.

Hurst, L. D., and N. G. C. Smith. 1999. Do essential genes evolve slowly? Curr. Biol. **9**:747–750.

Jordan, I. K., I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. **12**:962–968.

Lynch, M., and J. S. Conery. 2000. The evolution fate and consequences of duplicate genes. Science **290**:1151–1155.

Pearson, W. R., and D. J. Lipman. 1988. Improved tools for biological sequence analysis. Proc. Natl. Acad. Sci. USA **85**:2444–2448.

Steinmetz, L. M., C. Scharfe, A. M. Deutschbauer et al. (11 co-authors). 2002. Systematic screen for human disease genes in yeast. Nat. Genet. **31**:400–404.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. ClustalW-improving the sensitivity of progressive multiple sequence alignment through sequence weighting position–specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Wilson, A. C., S. S. Carlson, and T. J. White. 1977. Biochemical evolution. Annu. Rev. Biochem. **46**:573–639.

Winzeler, S. A., D. D. Shoemaker, A. Astromoff et al. (52 co-authors). 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. Science **285**:901–906.

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

Naruya Saitou, Associate Editor

Accepted December 24, 2002