Energy Constraints on the Evolution of Gene Expression

Andreas Wagner

Department of Biology, The University of New Mexico

I here estimate the energy cost of changes in gene expression for several thousand genes in the yeast *Saccharomyces cerevisiae*. A doubling of gene expression, as it occurs in a gene duplication event, is significantly selected against for all genes for which expression data is available. It carries a median selective disadvantage of $s > 10^{-5}$, several times greater than the selection coefficient $s = 1.47 \times 10^{-7}$ below which genetic drift dominates a mutant's fate. When considered separately, increases in messenger RNA expression or protein expression by more than a factor 2 also have significant energy costs for most genes. This means that the evolution of transcription and translation rates is not an evolutionarily neutral process. They are under active selection opposing them. My estimates are based on genome-scale information of gene expression in the yeast *S. cerevisiae* as well as information on the energy cost of biosynthesizing amino acids and nucleotides.

Introduction

Two major kinds of genetic change can affect rates at which genes are expressed. The first is mutation in regulatory regions that affects either transcription or translational efficiency. The second is gene duplication. If a gene duplication creates identical copies of a gene and its regulatory region, then the initial effect of the duplication is effectively a doubling in gene expression. Both kinds of genetic change play major roles in biological evolution. For example, a growing number of genome-scale expression studies reveal that substantial genetic variation in messenger RNA (mRNA) expression levels exists within populations, among populations, and among closely related species (Oleksiak, Churchill, and Crawford 2002; Townsend, Cavalieri, and Hartl 2003; Fay et al. 2004; Wittkopp, Haerum, and Clark 2004). Similarly, genome sequence analysis has shown that single-gene duplications occur at substantial rates in eukaryotic genomes (Lynch and Conery 2000; Gu et al. 2002). Between 30% and 50% of a eukaryotic genome's gene content consists of duplicated genes (Rubin et al. 2000; Conant and Wagner 2002). These observations underscore the evolutionary importance of single-gene duplication and the ensuing expression changes.

Increases in gene expression incur energy costs. The central question I pose here is whether these costs are substantial enough to affect the reproduction rate of organisms where rapid cell division is important for evolutionary persistence—most notably microbes. To be sure, rapid cell division in a nutrient-rich environment is only one among multiple factors influencing a microbe's success in surviving and reproducing in the wild. Other factors include surviving starvation, drastic temperature fluctuations, and osmotic shocks. Nonetheless, the evolutionary importance of rapid cell division is indicated by codon usage patterns that allow rapid protein synthesis when nutrients are abundant (Sharp and Li 1986; Akashi and Gojobori 2002).

Maximizing the energy available to cells for biosyntheses, growth, and division is essential for rapid cell division. This is vividly illustrated by recent work that analyzes how the input into a metabolic reaction network and the output produced by the network can affect cell growth (Ibarra, Edwards, and Palsson 2002; Segre, Vitkup, and Church 2002). However, we currently do not know whether gene expression changes in most genes would affect a cell's energy budget substantially enough to change cell division rates. For one thing, the question is almost impossible to address experimentally. One reason is that an experimental manipulation of gene expression may well carry energy costs, but the changed concentration of a gene product may also affect important biological processes. The two effects are very difficult to disentangle. A second reason is that tiny differences in growth rates, of the order of 10^{-7}, much smaller than can be measured in the laboratory, can affect the fate of mutants in microbes with large population sizes (Hartl and Clark 1997).

I here estimate the cost of changing RNA and protein expression relative to the total energy cost of gene expression in the eukaryotic microbe *Saccharomyces cerevisiae*. In order to do so, I use information on the energy cost (in activated phosphate bonds, ~P) of synthesizing the nucleotide building blocks of mRNA and the amino acid building blocks of proteins, as well as genome-scale information on the abundances and decay rates of mRNA and proteins. I relate these cost estimates to a critical selection coefficient *s*, estimated from the amount of nucleotide polymorphisms in the closest wild relative of *S. cerevisiae* (Johnson et al. 2004). The fate of mutations that change the energy budget by an amount smaller than this critical *s* is dominated by genetic drift. I show that the doubling of both protein and RNA expression, as might be caused by a gene duplication, carries energy costs much higher than *s* for all yeast genes for which expression data is available. Most assumptions I make in these estimates are conservative, such that improved data will likely show the actual energy costs of gene expression to be higher than what my results suggest. If true in general for microbes, this means that substantial changes in mRNA and protein synthesis rates, as well as gene duplications, can only go to fixation in a population when they provide an advantage sufficiently great to override these costs.

Results

Estimation of Energy Costs and the "Median Gene"

I here estimate the cost *s* of expressing any one yeast gene at experimentally observed levels as a fraction of the total energy consumed in yeast gene expression. The energy currency I use is the activated (high-energy) phosphate

Key words: gene duplication, evolution, gene expression.

E-mail: wagnera@unm.edu.

Mol. Biol. Evol. 22(6):1365–1374. 2005
doi:10.1093/molbev/msi126
Advance Access publication March 9, 2005

bond (~P). This expression cost can be partitioned into two components. The first of them is the energy needed to synthesize the ribonucleotide building blocks of a gene's mRNA and the amino acids for the gene's protein product. I determine this cost for growth on a minimal medium with glucose as sole carbon source under respiratory and fermentative conditions, using the DNA sequence of yeast genes and known precursor biosynthesis pathways (Neidhardt, Ingraham, and Schaechter 1990; Stryer 1995). The building block cost for mRNA includes the nucleotides for an untranslated region and a polyA tail with empirically observed lengths (Hurowitz and Brown 2004). The second cost component is the polymerization cost needed to make an mRNA molecule and a protein from their respective building blocks. This cost is small for mRNA because the RNA precursors are already activated molecules. In contrast, it is large for proteins because of the cost of charging transfer RNA (tRNAs) with amino acids and because of substantial elongation costs during translation (Stryer 1995). These two cost components are combined in the following way to estimate the energy cost per unit time of expressing a yeast gene at observed rates. If s_R is the rate at which a mRNA molecule is synthesized per second, R is the number of molecules of this mRNA in the cell, and d_R is the decay constant of this molecule (in s^{-1}), then the temporal change in the mRNA concentration R, dR/dt, is given by $dR/dt = s_R - d_R R$. In steady state, that is, when $dR/dt = 0$, one can calculate the synthesis rate s_R from the mRNA concentration and the decay constant d_S. Both have been measured for thousands of yeast genes (Wang et al. 2002; Arava et al. 2003), so that one can establish their distribution. If C_R is the energy cost of synthesizing one specific mRNA molecule, then the per-second synthesis cost of synthesizing this mRNA to maintain the experimentally observed level in the cell is simply $s_R C_R = R d_R C_R$ (~P s^{-1}). For proteins, with completely analogous notation, the synthesis cost calculates as $s_P C_P = P d_P C_P$ (~P s^{-1}). The overall cost is then the sum of the two, $R d_R C_R + P d_P C_P$ (~P s^{-1}). The expression cost for all protein-coding genes is simply the sum of the expression costs for individual genes.

For the purpose of illustration, the following simple calculation applies these considerations to a hypothetical average (median) gene. The listed numerical values are calculated from published information on yeast genes and their expression (Wang et al. 2002; Arava et al. 2003; Ghaemmaghami et al. 2003; Huh et al. 2003). The median length of a yeast RNA molecule is 1,474 nucleotides, and the median cost of precursor synthesis per nucleotide (derived from the base composition of yeast-coding regions) is 49.3 ~P. With a median mRNA abundance of R = 1.2 mRNA molecules per cell and a median mRNA decay constant of $d_R = 5.6 \times 10^{-4}$ s^{-1}, the mRNA synthesis costs calculates as $49.3 \times 1{,}474 \times 1.2 \times (5.6 \times 10^{-4}) = 48.8$ ~P per second and cell. This is a fraction $48.8/1.34 \times 10^7 = 3.6 \times 10^{-6}$ of the total RNA synthesis cost per second. The median length of a yeast protein is 385 amino acids, with a combined biosynthesis and polymerization cost of 30.3 ~P per amino acid. The median abundance is 2,460 protein molecules per cell. No currently available data allows a meaningful estimate of the median protein half-life, but a protein of an intermediate half-life (see below) of 10 h (decay constant $d_P = 1.92 \times 10^{-5}$ s^{-1}) yields an overall synthesis cost of $30.3 \times 385 \times 2{,}460 \times (1.92 \times 10^{-5}) = 551$ ~P s^{-1}.

The Relative Energy Investment into RNA and Protein

Quantitative information on all the cost components I listed above (R, d_R, C_R, P, d_P, C_P) can be derived for thousands of genes from publicly available information, with one exception: no unbiased, systematic, and reliable measurements are available for the decay constants (d_P) or, equivalently, the half-lives ($\tau_{1/2} = (\log_e 2)/d_p$) of most proteins. Using the N-end rule, yeast proteins can be engineered to have a half-life between 2 min and greater than 20 h (Varshavsky 1996). However, large-scale estimates of half-lives exist only for proteins isolated from two-dimensional electrophoresis experiments, which are high-abundance proteins with long half-lives (Gygi et al. 1999; Pratt et al. 2002).

I pursued two approaches to get meaningful expression cost estimates despite this lack of systematic information. I call the first the ribosomal occupancy (RO) approach. It uses estimated protein synthesis rates s_P based on the empirically observed number of ribosomes attached to a mRNA (Arava et al. 2003). From this estimate and from known protein abundance, one can infer d_P and use it to estimate costs in the manner outlined above. However, these observed protein synthesis rates have not yet been independently confirmed and have to be taken with much caution. Thus, I also pursue a second approach based on the HL of abundant proteins. It rests on the observation that the vast majority of protein synthesis activity in any cell goes towards high-abundance proteins, because the distribution of protein abundances is highly skewed. For instance, if one considers as highly abundant those proteins with a higher than median abundance (2,460 copies per cell), then more than 95% of protein molecules in a cell fall into the high-abundance class (Ghaemmaghami et al. 2003). In other words, if one were to choose a protein molecule from a cell at random, there would be a greater than 95% chance that the chosen protein exists in more than 2,460 copies per cell. Proteins above the median abundance can be readily extracted from two-dimensional electrophoretic gels (Gygi et al. 1999) and their half-life determined. Such proteins overwhelmingly tend to have long half-lives (Gygi et al. 1999; Pratt et al. 2002). These considerations suggest that the total energy cost of protein synthesis is approximated by the cost of the synthesis of proteins of long half-lives. In a systematic effort to estimate the half-lives of more than 50 yeast proteins isolated from two-dimensional gels, Pratt et al. (2002) estimated a mean decay constant of 0.022 h^{-1} (6.1×10^{-6} s^{-1}), which corresponds to a half-life of 31.51 h. Taken together, these observations suggest that the total energy cost of protein synthesis will be well approximated if one assumes that all proteins have the average half-life typical of that of abundant proteins. This assumption is the basis of the second, "HL" (half-life) approach. Both the RO and half-life estimates of the total amount of energy going into protein synthesis are within one order of magnitude of 10^7 ~P s^{-1}.

The fact that proteins occur at manifold greater abundances in cells than their respective mRNAs (Ghaemmaghami et al. 2003) suggests that a much greater amount of energy

goes into their production than into that of mRNA. However, there are also several factors that make mRNA production more expensive relative to that of proteins. First, the average combined synthesis and polymerization costs are more than 50% higher for nucleotides (49.3 ~P per nucleotide) than for amino acids (30.3 ~P per amino acid). Second, mRNAs consist of more than three times the number of monomers than the proteins they encode. Third, proteins, and especially proteins of greater than median abundance tend to have much longer half-lives than mRNA. The longer a molecule's half-life, the smaller the necessary synthesis rate s_P (and thus the energy needed) to sustain a given steady-state level. These factors will conspire to raise the mRNA synthesis costs towards those of proteins. According to the available data, total mRNA synthesis cost (6.69×10^5 ~P s^{-1}) is, however, still smaller than the total protein synthesis costs (RO: 1.55×10^7 ~P s^{-1}; HL: 6.22×10^6 ~P s^{-1}).

Estimation of Selection Coefficients

To estimate the impact of expressing any one gene on a cell's energy budget, one would ideally want to know the total energy consumed per growing cell and second. Although this total energy is unknown, it is clear that gene expression accounts for the majority of it. First, 51.3% of yeast biomass consists of RNA and proteins (Forster et al. 2003). (Most of the remaining biomass [>39%] consists of various polymers of glucose and its relatives, specifically glycogen, mannan, and glucan. Only 2.9% and 0.4% of the biomass consists of lipids and DNA, respectively.) In addition, 76.6% of the total adenosine triphosphate (ATP) cost of polymerization is invested into RNA and protein polymerization and most of the remainder (21.8%) into carbohydrate biosynthesis (Forster et al. 2003). Taken together, these considerations suggest that the estimated fractional energy cost of changing a gene's expression is no more than a factor two higher than the actual energy cost. I take this uncertainty of at most a factor two into account below.

The expression of any one gene consumes only a small fraction of a cell's energy budget. But how small must this fraction be so as to be "invisible" to natural selection? In a diploid organism, the magnitude of a selection coefficient s below which genetic drift influences the fate of a genotype more strongly than natural selection (the "critical" selection coefficient) is $s = 4/N_e$ (Hartl and Clark 1997). Here N_e is the effective size of a population, and s the selective disadvantage (selection coefficient) caused by the energy cost of an increase in gene expression. The effective population size can be estimated from the nucleotide diversity π at synonymous sites, because the expected nucleotide diversity for such neutral sites is equal to $\pi = N_e\mu$, where μ is the mutation rate per nucleotide site. In sum, one has $s = 4/N_e = 4\mu/\pi$.

Saccharomyces cerevisiae has long been associated with humans, even before being used as laboratory organisms (Mortimer 2000). Its recent population structure may thus not reflect its evolutionary history. For this reason, I chose to use diversity estimates from *S. cerevisiae*'s closest wild relative, *Saccharomyces paradoxus*, whose sequence divergence to *S. cerevisiae* is approximately 0.11 nucleotide substitutions per nucleotide site in genic regions (Kellis et al. 2003). For *S. paradoxus* $\pi = 0.003$ (Johnson et al. 2004). Together with an estimated mutation rate of $\mu = 2.2 \times 10^{-10}$ (Drake et al. 1998) this yields an effective population size estimate of $N_e = 0.003/(2.2 \times 10^{-10}) = 1.36 \times 10^7$. From this estimate follows a critical selection coefficient of $s = 4\mu/\pi = 4 \times (2.2 \times 10^{-10})/0.003 = 2.93 \times 10^{-7}$. To take into account that the fractional energy costs I calculate here may slightly overestimate the actual fractional cost (see above), I incorporate a factor one-half into this estimate, yielding a critical s of $s < (2.93 \times 10^{-7})/2 = 1.47 \times 10^{-7}$ ($\log_{10}(s) = -6.83$), below which genetic drift dominates a genotype's evolution.

The Energy Cost of Gene Duplication Is Significant

Figure 1*a* shows a histogram of the fractional energy costs (energy cost as a fraction of the total cost of gene expression, under respiratory conditions) associated with the simultaneous doubling of RNA and protein expression, as would occur in a typical single-gene duplication. It is crucial to appreciate the logarithmic scale of the x axis, where every unit change indicates a change by a factor 10 in the fractional expression cost. The figure is based on the RO approach. Figure 1*b* shows the same distribution, but for the HL approach. The conclusion from both panels is the same: For all genes considered here, the fractional expression cost is much greater than the critical selection coefficient of $s = 1.47 \times 10^{-7}$ below which the fate of a gene duplicate would be dominated by drift in all three scenarios examined. It is important to realize that these are not genes of unusually high abundance. Their products span more than three orders of magnitude in mRNA abundances (0.1 to 131 copies per cell) and more than four orders of magnitude in protein abundances (50 to more than 10^6 copies per cell). Essentially, the same results as shown for respiratory conditions in figure 1 hold for fermentative conditions (median $s = -4.28/-4.75$; minimum $s = -5.82/-6.17$ for RO/HL, respectively). The shape of the distributions approximates a normal distribution and indicates that fractional expression costs are log-normally distributed. The 10 genes with the highest fractional expression cost ($0.0028 < s < 0.01$) for the RO data include seven genes encoding glycolytic enzymes, a ribosomal protein-coding gene (RPS8A), and a gene encoding the translation elongation factor eEF3.

These results imply that duplicates of the 4,346 yeast genes for which the necessary data are available would be subject to negative selection sufficiently strong to influence their fate, if they were expressed at the same level as the originals. At the same time, the data show that the vast majority of genes carry an expression cost whose effect would be too small to detect in a laboratory evolution experiment. For example, the median logarithm of the fractional cost in figure 1*a* is −4.33, corresponding to a selection coefficient of $s = 4.68 \times 10^{-5}$. In a large yeast cell population consisting of equal numbers of cells with a wild-type genotype and a genotype that grows more slowly by a factor $(1 - s)$, it would take $t_{1/2} = 1/s$ or more than 21,000 generations to halve the population frequency of the more slowly growing genotype.

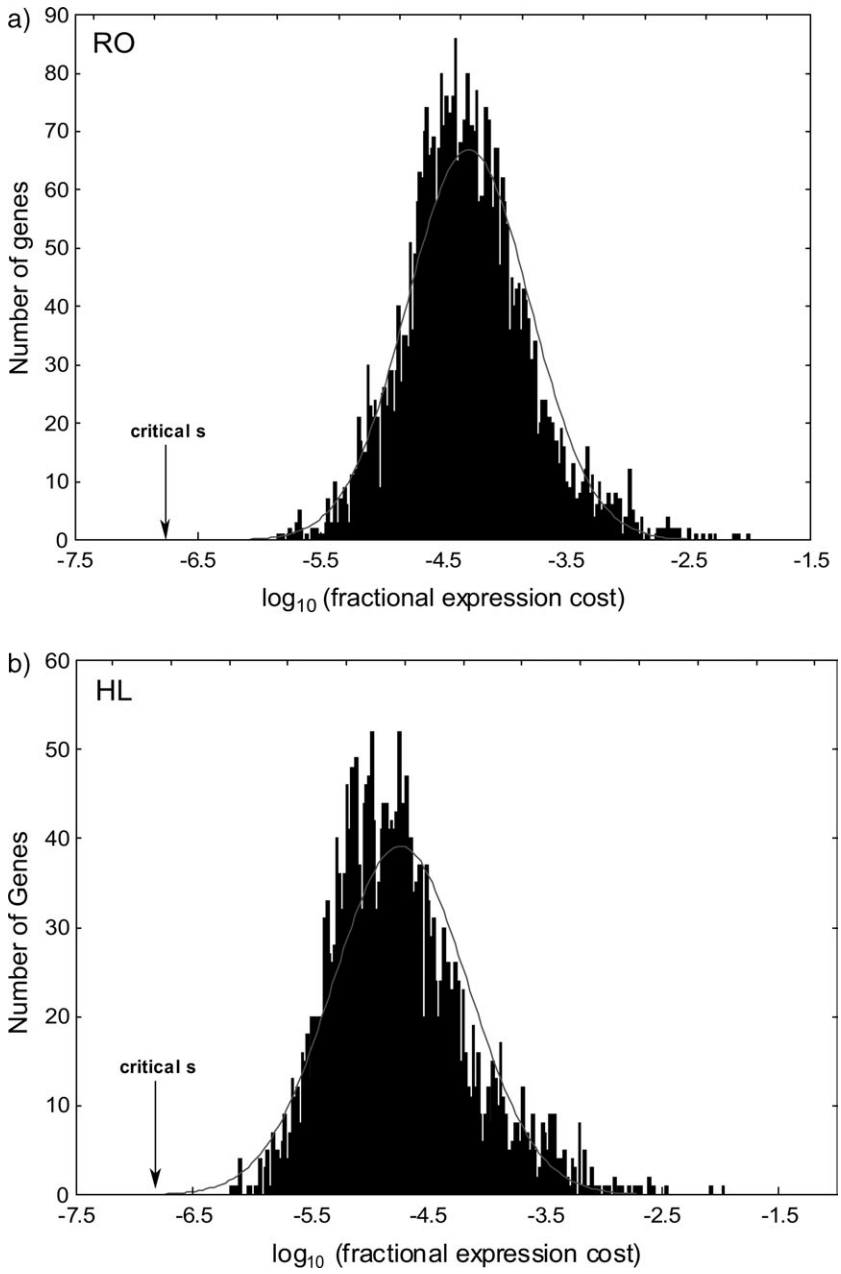

FIG. 1.—Distribution of the fractional energy cost of simultaneously doubling messenger RNA (mRNA) and protein expression of a gene. (*a*) Ribosomal occupancy approximation. (*b*) Long half-life approximation**.** Note the logarithmic scale on the horizontal axis. The arrow points to the fractional cost below which a change is effectively neutral. The black curve shows a fit to a normal distribution.

RNA and Protein Synthesis Rates Can Change Neutrally Only by Small Amounts

While a gene duplication can double synthesis of both RNA and protein products of any one gene, most regulatory mutations may change either RNA or protein synthesis, but not both. I thus asked in two complementary ways how much RNA or protein synthesis could change without substantially affecting a cell's energy budget. First, I determined the distribution of selection coefficients associated with a doubling of mRNA expression levels. The results are very similar to those obtained above for a joint doubling of mRNA and protein levels. That is, in the vast majority of genes, a doubling of mRNA expression incurs energy costs sufficiently large for natural selection to counteract. Specifically, under respiratory conditions the median selection coefficient counteracting a doubling in mRNA expression is greater than 10^{-6} ($1.97 \times 10^{-6}/2.9 \times 10^{-6}$ for RO/HL) and more than one order of magnitude greater than the critical selection coefficient of $s = 1.47 \times 10^{-7}$. For more than 99% of genes, the selection coefficient associated with a doubling of mRNA synthesis is greater than the critical selection coefficient (4,319 of 4,346 genes for RO, 4,340 of 4,346 genes for LH). Second, I asked for each gene by what factor mRNA expression can maximally increase such

that the associated energy cost is less than $s = 1.47 \times 10^{-7}$ and determined the distribution of this "neutral factorial change" (fig. 2a,b). Again, as a rule with some exceptions, it is small: its median is less than 0.1, that is, the mRNA synthesis rate can change by less than 10% without incurring significant energy costs (median/maximum 0.07/3.4 for RO, 0.05/2.3 for HL).

Similar to this analysis for mRNA gene expression, I determined the distribution of energy costs associated with changing protein synthesis. The median selection coefficients associated with doubling protein synthesis are greater than for mRNA expression change (3.89×10^{-5} for RO; 9.15×10^{-6} for HL). The smallest selection coefficients are greater than the critical $s = 1.47 \times 10^{-7}$ ($s = 5.82 \times 10^{-7}$/ 1.75×10^{-7} for RO/LH). The neutral factorial change in protein synthesis is also smaller than for mRNA, with a median smaller than 0.02 and a maximum smaller than one (median/maximum: 0.0038/0.25 for RO; 0.016/0.84 for HL). This means that a change in protein synthesis rate exceeding 2% incurs significant energy costs for more than half of all genes. In other words, protein synthesis rates are energetically substantially more constrained than mRNA synthesis rates. This is also evident from figure 2c, which plots the maximally permissible neutral change for mRNA synthesis versus that for protein synthesis. The straight line indicates equality. The figure shows, as one might expect, that the neutral factorial changes in mRNA and protein expression are correlated (Spearman $s = 0.53$; $P < 10^{-10}$). However, there is also substantial scatter in this correlation ($r^2 = 0.07$). This scatter indicates substantial variation in degradation rates (s_R, s_P), gene lengths, and synthesis costs of building blocks, all of which also influence the neutral factorial change in synthesis rates. mRNA and protein synthesis rates are similarly constrained also under fermentative conditions (median neutral factorial change in mRNA expression 0.09/0.06 for RO/HL; in protein expression 0.0033/0.013 for RO/HL).

Discussion

The results above show that increases in mRNA and protein synthesis in yeast are constrained through the energy costs they incur. The simultaneous doubling of mRNA and protein synthesis, which accompanies many gene duplications, causes a reduction in a cell's energy budget (and thus reduced growth) that is at least severalfold higher than the critical selection coefficient $s = 1.47 \times 10^{-7}$, below which a genotype's evolution is dominated by genetic drift. Even when considered separately, mRNA and protein synthesis can increase on average by no more than 10% without causing significant energy costs. This holds under both respiratory and fermentative conditions. It may seem surprising until one considers that even single amino acid substitutions in a protein may cause significant changes in energy costs due to the different biosynthesis costs of different amino acids (Akashi and Gojobori 2002).

Caveats

Because these results rely on genome scale but still incomplete information, they require estimates of some important quantities. Importantly, most of these estimates are conservative: They provide lower bounds on energy costs. Actual costs may well be higher for many genes.

The quantity least well characterized on a genome-wide scale is protein half-life. With regard to it, I take two complementary approaches. The first approach indirectly estimates protein synthesis rates from protein abundances and experimentally observed ribosomal occupancies (Arava et al. 2003). The second approach takes advantage of the observation that overall energy investment into protein synthesis is overwhelmingly dominated by proteins with high abundance and assumes that all proteins have a turnover rate identical to the average rate of abundant proteins (Pratt et al. 2002). Both approaches yield very similar results for the distribution of fractional energy costs of gene expression—the main focus of this work. However, the second approach may greatly underestimate the cost of expressing proteins of intermediate and low abundance. The reason is that many lowly expressed proteins may have a half-life much shorter than the half-life typical of highly expressed proteins. In consequence, their expression at observed levels (Ghaemmaghami et al. 2003) will consume more energy than estimated here. In other words, changing the expression level of these proteins will be even more costly than what I estimated.

A second source of uncertainty arises from missing expression information for some 40% of yeast genes. I here assumed that the distribution of expression costs for these genes is similar to that of the genes whose abundances have been measured. However, many genes may not be expressed at all in any one experimental condition. If so, my estimate of the total energy investment into gene expression will be an overestimate. Thus, the fractional energy invested into the expression of any one gene will be an underestimate. That is, again, gene expression will be even more costly than what is shown here by me.

A third caveat pertains to gene duplications. I have neglected the energy cost caused by the additional DNA introduced through a single-gene duplication. Is this energy cost comparable to that incurred by the additional gene expression? No. It is much smaller. This is evident from the fact alone that DNA constitutes only 0.4% of yeast biomass, whereas proteins and RNA constitute more than 50% (Forster et al. 2003). (Replication time delays caused by additional DNA may be a problem in prokaryotes with one origin of replication but less so in eukaryotes which have multiple origins of replication.)

The Influence of Population Parameters

I purposedly separated the analysis of gene expression costs in that of a fractional energy cost and that of a critical selective coefficient s. While the total gene expression cost in any one physiological state may vary little among individuals in a species, critical selection coefficients—coefficients s below which a mutation's fate is dominated by genetic drift—depend on the (effective) population size N_e of a population. (In large populations, selection against a mutation with increased gene expression may be weak [small s] and still eliminate the mutation.) Estimates of N_e in turn, rely on other parameters such as the number

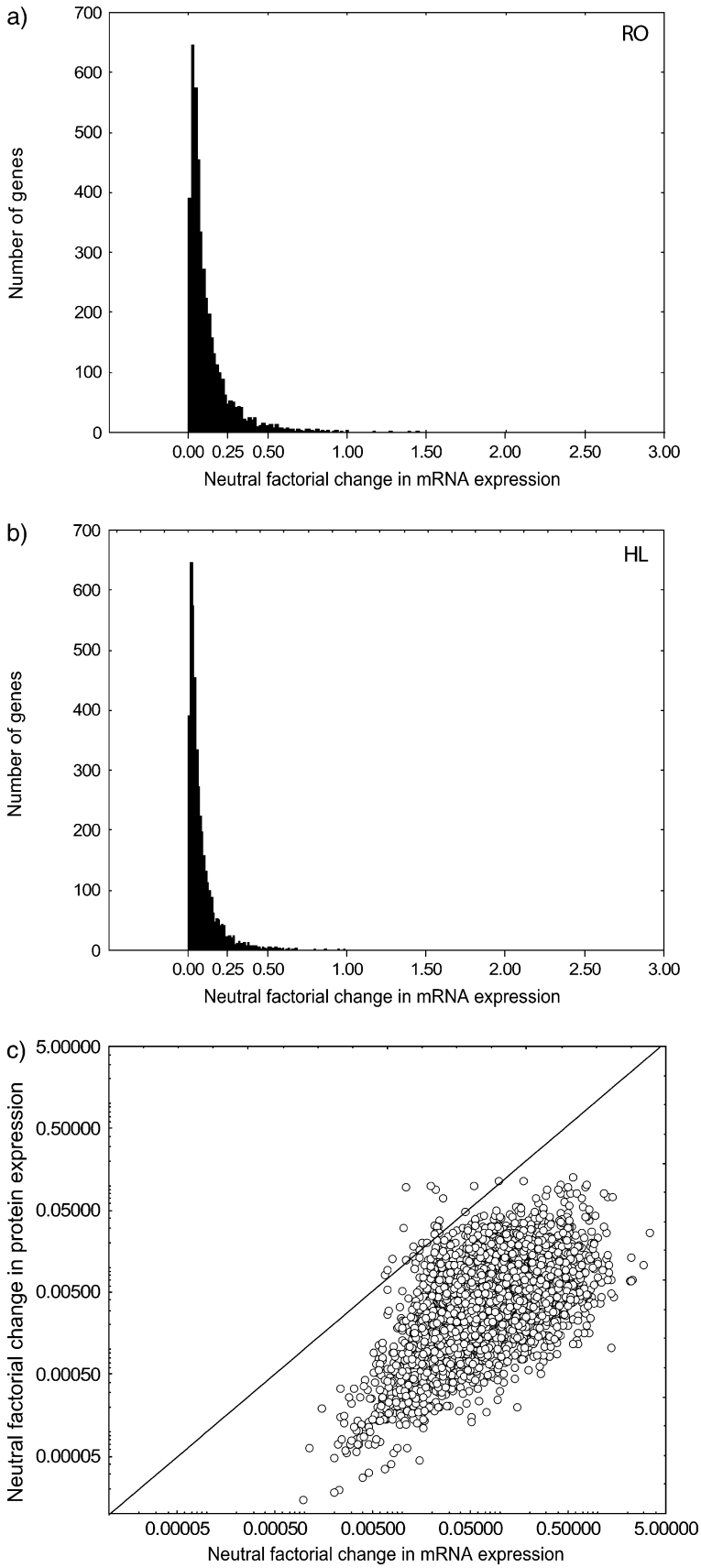

FIG. 2.—Fractional changes in mRNA expression that are effectively neutral in (*a*) the ribosomal occupancy (RO) and (*b*) the long half-life (LH) estimate of protein synthesis costs. If C is the cost of synthesizing a gene product (mRNA), if T is the total (genome-wide) RNA and protein synthesis cost

of nucleotide polymorphisms and the mutation rate. Several caveats follow. First, nucleotide polymorphisms may be under the influence of selection. I minimized the likely influence of selection by using only data from silent (synonymous) nucleotide polymorphisms. If, however, some synonymous sites were under negative or positive selection, the nucleotide diversity π at synonymous sites would underestimate the number of neutral polymorphisms to be expected. This would make my estimate of the critical s, $s = 4\ \mu/\pi$, too large and my conclusions conservative because π for neutral sites should actually be greater than the available estimate. (The opposite would hold for balancing selection, but such selection is rather unusual even for amino acid replacement sites.) Second, the mutation rate estimate μ is derived from laboratory populations. Such estimates have been criticized as too high for other microbes (Ochman, Elwyn, and Moran 1999). If so, then s should be even smaller than what I estimated and gene expression increases would be even costlier than it appears in my analysis.

A third consideration is that the relation $\pi = N_e\mu$ I use here is based on the assumption of random mating, whereas yeast undergoes a significant amount of selfing. Pertinent work (Nordborg and Donnelly 1997) shows that in selfing organisms evolution by random genetic drift is accelerated, that is, the effective population size is smaller than in randomly mating organisms. This effect, however, is rather moderate. Specifically, even for exclusively selfing organisms—an extreme case—N_e is reduced only by a factor of two, meaning that s would have to be raised by the same amount.

Finally, my estimate of N_e is based on synonymous nucleotide polymorphism data in *S. paradoxus*, the closest wild relative of *S. cerevisiae*. The effective population size $N_e = 1.36 \times 10^7$ lies between that of prokaryotic microbes like *Escherichia coli* (2×10^8; Hartl et al. 1994) and higher eukaryotes like *Drosophila* (5×10^6; Ayala et al. 1993) but is subject to revision based on improved population data. Such a revision may well lead to an increase in N_e, considering that the current estimate is based on a local population sample in an area only 10 km^2 in size (Johnson et al. 2004). (The population size relevant for estimating s is the global population size, which may be much larger.) If so, even smaller gene expression costs than I suggest here may have evolutionary consequences.

Some of my results are highly robust to changes in N_e. Specifically, gene duplication (or a doubling of either mRNA synthesis or protein synthesis) would still incur significant energy costs for most genes, even if the critical s changed severalfold (figs. 1 and 2). In contrast, the neutral factorial change in mRNA and protein expression depends linearly on this selection coefficient. A halving of the effective population size would lead to a doubling of the effectively neutral mRNA and protein synthesis change. In other words, even if other organisms have an energy cost distribution of gene expression similar to that of yeast, their mRNA and protein synthesis rates may be under different constraints. They may be less constrained, for example, in pathogens which may suffer periodical bottlenecks—and thus reduced N_e—before infecting a new host.

Higher Organisms

All of the above applies only to microbes. In higher organisms, energetic constraints on gene expression may be of minor evolutionary importance because other components of fitness, especially behavioral components, dominate. To be sure, changes in gene expression could affect a cell's energy budget, which could affect cell division rates. Major changes in the timing of cell division, in turn, can affect important events in embryonic development and even lead to a reorganization of the embryonic body plan. However, most changes in gene expression would affect the timing of cell division by only a small amount because they consume only a small fraction of a cell's energy budget. Nonetheless, the reduced importance of energy constraints in higher organisms has one potentially important evolutionary consequence: newly arisen regulatory mutations that increase gene expression and gene duplication may be able to go to fixation more easily. It has been proposed that rates of single-gene duplication and other important events in genome evolution increase in higher organisms because of their lower effective population sizes (Lynch and Conery 2003). I speculate that a reduced importance of energy considerations in genome evolution contributes to this increased incidence of gene duplications.

How Can Gene Expression Change on an Evolutionary Timescale?

The observation that energy costs constrain the evolution of gene expression raises the question how substantial gene expression changes arise on evolutionary timescales. There are at least three scenarios, all of which may operate at the same time. The first of them involves changes in mRNA or protein half-life. An increase in the cellular concentration of a gene product can be achieved either by increasing its synthesis or by increasing its half-life. Whereas an increase in synthesis costs energy, a decrease in half-life does not. From this point of view, one would predict that changes in half-life would contribute importantly to changes in gene expression because half-lives are energetically less constrained than synthesis rates. (This argument neglects other costs of increasing half-lives, such as a reduced ability to regulate gene expression dynamically, in response to environmental changes. The importance of these costs is unknown.) Second, every large population experiences a substantial influx of regulatory mutations. Some of these mutations may increase synthesis of some gene products and decrease synthesis of others,

←

($\sim$P s^{-1}), and if s is the selection coefficient below which the fate of a mutant is dominated by genetic drift, then the maximal fractional change f_{max} in the synthesis whose fate is dominated by genetic drift is $f_{max} = s$T/C. It is the distribution of f_{max} that is plotted here. (*c*) The neutral factorial change in mRNA synthesis plotted against the neutral factorial change in protein expression (determined analogous to that for mRNA). Note the double-logarithmic scale, the smaller neutral factorial change in protein expression, and the considerable scatter.

such that the total energy consumption may remain unchanged. On a genome-wide scale, multiple mutations that compensate each other's effects on the overall energy balance may be an important mode of gene expression evolution. The third and perhaps most important mechanisms involves selection. For a mutation that causes a substantial increase in mRNA or protein synthesis to go to fixation, the mutation needs to have even greater fitness benefits to overcome this effect of selection opposing it. The benefits of increased expression are most obvious for both lowly and highly expressed genes. In lowly expressed genes, gene expression noise may cause a gene's expression level to dip below levels necessary for proper biological functions. Increased expression or gene duplication may eliminate these dips (Cook, Gerber, and Tapscott 1998). On the other extreme, increased expression of genes whose products are in particularly high demand can increase growth. This is obvious for enzyme-coding genes or transporters, where a higher concentration of gene product may permit higher metabolic flux, and thus an increased rat of biosynthesis or energy production. However, it seems to be quite a general phenomenon: In 11 different functional classes of genes, genes with a high codon usage bias—which indicates high expression—have more duplicates than other genes (Papp, Pal, and Hurst 2003). The trade-offs between costs and benefits of raised gene expression are uncharted territory and well worth exploring further.

In all of the above, it is worth remembering that the selective disadvantage incurred by even a doubling of expression is minute for most genes and would lead to elimination only in the course of thousands of generations. These small differences notwithstanding, the appropriate null model for the evolution of transcription and translation rates is not a neutral model, at least for microbes. It is a model where natural selection opposes increases in transcription and translation rates in individual genes on energetic grounds.

Methods

Metabolic Costs of Amino Acid and Nucleotide Synthesis

All nucleotides and amino acids are synthesized from a small number of metabolic precursors through biosynthetic pathways that are highly conserved among free-living organisms. These precursors are important intermediates in energy metabolism, such as pyruvate and 3-phosphoglycerate. The overall cost of making one amino acid or nucleotide is thus equal to the loss of energy that could have been be produced if the precursor had not been removed from energy metabolism (the "precursor cost") plus the cost of making the amino acid or nucleotide from the precursor (the "biosynthesis cost"). I am using the activated phosphate (~P) of ATP and its relatives as the energy unit. Under respiratory conditions, reducing equivalents in the reduced form of nicotinamide adenine dinucleotide and related molecules are quantitatively transformed into ATP at stoichiometries that vary depending on where in energy metabolism a reducing equivalent has been generated. For reducing equivalents generated in glycolysis and the trichloroacetic acid cycle, I use stoichiometries given in (Stryer 1995, p. 552). For other reducing equivalents I use an average stoichiometry of 2 ~P per one reducing equivalent. To give one example, the precursor cost of pyruvate under respiratory conditions is 12.5 ~P. The reason is that one molecule of glucose is converted into two molecules of pyruvate and 30 ~P. Producing one pyruvate from glucose yields 2.5 mol of ~P. Thus, removal of one pyruvate for the biosynthesis of an amino acid effectively costs (30/2) − 2.5 = 12.5 ~P. Under fermentative conditions, both precursor and biosynthesis costs change because reducing equivalents are no longer converted into ~P but instead unloaded onto terminal electron acceptors such as acetaldehyde. To estimate energy costs under fermentative conditions with acetaldehyde as electron acceptor, I thus count only reactions that directly produce or consume ~P. Absolute energy costs for many amino acids and nucleotides appear lower under fermentative conditions because of the greater inefficiency of energy metabolism under these conditions. The following are the precursors necessary for amino acid and nucleotide biosyntheses, along with their costs under respiratory/fermentative conditions: pyruvate (12.5/0), 3-phosphoglycerate (14.5/1), phosphoenolpyruvate (14.5/1), acetyl-CoA (10/0), oxaloacetate (13.5/1), α-ketoglutatatrate (7.5/0), ribose-5-phosphate (27/3), and erythrose-4-phosphate (27/3).

For the biosynthesis costs of amino acids and nucleotides from their precursors, I take advantage of the nearly universal conservation of biosynthetic pathways and use tabulated pathway and cost data from *E. coli* (Neidhardt, Ingraham, and Schaechter 1990), with a stoichiometry of 2 ~P (0 ~P) per reducing equivalent under respiratory (fermentative) conditions. Using the above precursor costs, the total energy cost of amino acids and nucleotide precursors under respiratory/fermentative conditions calculates as follows: alanine (14.5/2), arginine (20.5, 13), asparagine (18.5/6), aspartate (15.5/3), cysteine (26.5/13), glutamate (9.5/2), glutamine (10.5/3), glycine (14.5/1), histidine (29/5), isoleucine (38/14), leucine (37/4), lysine (36/12), methionine (36.5/24), phenylalanine (61/10), proline (14.5/7), serine (14.5/1), threonine (21.5/9), tryptophan (75.5/14), tyrosine (59/8), valine (29/4), ATP (48.5/11), guanosine triphosphate (48.5/11), uridine triphosphate (51.5/15), cytidine triphosphate (49.5/13). I note parenthetically that the results reported here would be essentially unchanged (data not shown) if one took the much simpler approach of using cost data such as that reported in Neidhardt, Ingraham, and Schaechter (1990) directly, only considering the number of activated phosphate bonds expended in making a building block and ignoring the contribution of diverting a precursor from energy metabolism.

Synthesis Costs per mRNA Molecule and Protein

The nucleotide building blocks of mRNA are already activated molecules, and the synthesis cost of an mRNA molecule is thus well approximated by the synthesis costs of its constituent nucleotides. The synthesis cost of the translated region of the mRNA can be determined by adding the contributions of individual nucleotides from the gene's protein-coding region (obtained from the Saccharomyces Genome Database, http://www.yeastgenome.org/).

The second component is the energy cost of the untranslated region and the polyA tail. Although both vary to some extent among genes, a recent genome-scale study (Hurowitz and Brown 2004) showed that their respective lengths are well approximated by 256 and 60 nucleotides for most genes. Because the exact composition of the untranslated region is unknown, I used the average nucleotide composition of yeast coding regions (A:G:C:U = 0.327:0.204:0.192:0.277) as a substitute to estimate its energy cost. The resulting median cost for the mRNA expression of yeast genes is 49.3 (12.3) ~P under respiratory (fermentative) conditions. It varies among yeast genes between a minimum of 49.1 (11.9) ~P and a maximum of 49.7 (12.8) ~P.

As opposed to mRNA synthesis, protein synthesis carries substantial costs in addition to amino acid biosynthesis. The major additional cost components are 2 ~P for the charging of tRNAs with amino acids, 2 ~P for translation initiation, 2 ~P for each translocation of the ribosome along the mRNA during elongation, and 1 ~P for termination (Moldave 1985). These costs need to be added to the biosynthesis costs of amino acids. They yield a median cost per amino acid for the protein expression of yeast genes of 30.3 (10.6) ~P under respiratory (fermentative) conditions. Among yeast genes, this per-amino-acid-cost varies between 22.1 (7.4) and 40.5 (15.1) ~P, a range that is much narrower than that between the "cheapest" amino acid and costliest amino acids. In these calculations, I excluded some minor cost components, such as energy costs of proofreading and the cost of ribosomal scanning of the mRNA to find the start codon because the precise energy requirements for these processes are unknown (Moldave 1985). To the extent that these processes affect most genes to an equal extent, I note that they will not affect my conclusions because I am focusing not on the absolute amount of energy invested in the expression of any one gene but on the proportion of this energy relative to the total gene expression cost.

mRNA and Protein Synthesis Costs per Unit Time

All calculations assume prolonged exponential cell growth, where mRNA and protein synthesis rate have reached a steady-state characteristic for this growth phase and where new protein and mRNA synthesis is fed by newly synthesized monomers and not through salvage pathways using recycled components. To estimate mRNA synthesis cost I used published data on mRNA abundances R and decay rates d_R (Wang et al. 2002, Arava et al. 2003; http://genome-www.stanford.edu/yeast_translation/supplement.shtml) for 4,379 yeast genes that are expressed in exponentially growing cells. The respective data include mRNAs with a wide range of abundances (0.1 to 130 mRNA copies per cell) and decay rates (2.46×10^{-5} to 4.54×10^{-3} s^{-1}). For each of these genes, I determined the expression cost per second as Rd_RC_R, where C_R is the energy cost (in ~P) of producing the mRNA of the gene. The total cost of producing the mRNAs of these genes is obtained by adding up the individual cost. This cost, however, is still an underestimate of the total RNA expression cost of a cell. First, the available data does not contain information about the expression state of all 6,300 genes. Some of the remaining genes may not have been expressed under the experimental conditions, whereas the expression of others may not have been detected for technical reasons. In addition, mRNA accounts only for 5% (1/20th) of the total RNA of a cell (Ju, Morrow, and Warner 1990), the rest being expressed from RNA-coding genes such as those for tRNAs and ribosomal RNAs. I conservatively assumed that the expression ranges and decay rates of the remaining genes have a similar distribution to those in the available data and extrapolated to the total cost by multiplying with a factor ($20 \times 6{,}300/4{,}379$) = 28.77.

To estimate protein synthesis cost, I took two approaches necessitated by the absence of direct information about most protein HLs. The RO approach uses data on observed mRNA abundances and number of ribosomes per mRNA for 5,670 yeast genes to estimate protein synthesis rates (Arava et al. 2003). The second, LH approach uses information on the abundance of 3,570 yeast proteins in exponentially growing cells (Ghaemmaghami et al. 2003; Huh et al. 2003; http://yeastgfp.ucsf.edu) and assumes that all proteins have a decay constant equal to that of the average decay constant of long-lived proteins (6.1×10^{-6} s^{-1}; Pratt et al. 2002). (This approach will underestimate the synthesis cost for short-lived proteins, which means that the fractional cost of changing their expression will be even higher than what I estimate here.) Analogous to mRNA synthesis costs, protein synthesis costs are again calculated as Pd_PC_P, and I extrapolated to the total synthesis costs by assuming that the distribution of synthesis costs of proteins for which empirical data is available is representative of all proteins. All data I use here have been obtained in derivatives of the yeast strain S288C.

Gene expression costs are greatest in a minimal medium where a cell needs to synthesize all amino acids and nucleotides. My calculations of expression cost are based on such a medium, but mRNA and protein abundance distributions as well as information on decay rates are available only for more complex media. Because cell division times vary by less than a factor two in these two kinds of media (Sherman 1991), the total energy invested into gene expression are probably no more variable than that. However, an assignment of energy costs to individual genes can currently not be made for this reason. This is why I restrict myself to characterizing the distribution of energy cost.

Acknowledgments

I would like to thank the NIH for its support through grant GM63882 to the Department of Biology at the University of New Mexico as well as the Santa Fe Institute and the iHES for their continued support.

Literature Cited

Akashi, H., and T. Gojobori. 2002. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proc. Natl. Acad. Sci. USA **99**:3695–3700.

Arava, Y., Y. L. Wang, J. D. Storey, C. L. Liu, P. O. Brown, and D. Herschlag. 2003. Genome-wide analysis of mRNA

translation profiles in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA **100**:3889–3894.

Ayala, F. J., and D. L. Hartl. 1993. Molecular drift of the bride of sevenless (boss) gene in Drosophila. Mol. Biol. Evol. **10**:1030–1040.

Conant, G. C., and A. Wagner. 2002. GenomeHistory: a software tool and its applications to fully sequenced genomes. Nucleic Acids Res. **30**:1–10.

Cook, D. L., L. N. Gerber, and S. J. Tapscott. 1998. Modeling stochastic gene expression: implications for haploinsufficiency. Proc. Natl. Acad. Sci. USA **95**:15641–15646.

Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow. 1998. Rates of spontaneous mutation. Genetics **148**: 1667–1686.

Fay, J. C., H. L. McCullough, P. D. Sniegowski, and M. B. Eisen. 2004. Population genetic variation in gene expression is associated with phenotypic variation in Saccharomyces cerevisiae. Genome Biol. **5**:R26.

Forster, J., I. Famili, P. Fu, B. Palsson, and J. Nielsen. 2003. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. Genome Res. **13**:244–253.

Ghaemmaghami, S., W. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman. 2003. Global analysis of protein expression in yeast. Nature **425**:737–741.

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

Gygi, S. P., Y. Rochon, B. R. Franza, and R. Aebersold. 1999. Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. **19**:1720–1730.

Hartl, D. L., E. N. Moriyama, and S. A. Sawyer. 1994. Selection intensity for codon bias. Genetics **138**:227–234.

Hartl, D. L., and A. G. Clark. 1997. Principles of population genetics. Sinauer Associates, Sunderland, Mass.

Huh, W. K., J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, and E. K. O'Shea. 2003. Global analysis of protein localization in budding yeast. Nature **425**:686–691.

Hurowitz, E. H., and P. O. Brown. 2004. Genome-wide analysis of mRNA lengths in Saccharomyces cerevisiae. Genome Biol. **5**:R2.

Ibarra, R. U., J. S. Edwards, and B. O. Palsson. 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature **420**:186–189.

Johnson, L. J., V. Koufopanou, M. R. Goddard, R. Hetherington, S. M. Schafer, and A. Burt. 2004. Population genetics of the wild yeast Saccharomyces paradoxus. Genetics **166**:43–52.

Ju, Q. D., B. E. Morrow, and J. R. Warner. 1990. Reb1, a yeast DNA-binding protein with many targets, is essential for cell-growth and bears some resemblance to the oncogene Myb. Mol. Cell. Biol. **10**:5226–5234.

Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. 2003. Sequencing and comparison of yeast genes to identify genes and regulatory elements. Nature **423**:241–254.

Lynch, M., and J. Conery. 2003. The origins of genome complexity. Science **302**:1151–1155.

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

Moldave, K. 1985. Eeukaryotic protein synthesis. Annu. Rev. Biochem. **54**:1109–1149.

Mortimer, R. K. 2000. Evolution and variation of the yeast (Saccharomyces) genome. Genome Res. **10**:403–409.

Neidhardt, F. C., J. Ingraham, and M. Schaechter. 1990. Physiology of the bacterial cell. Sinauer Associates, Sunderland, Mass.

Nordborg, M., and P. Donnelly. 1997. The coalescent process with selfing. Genetics **146**:1185–1195.

Ochman, H., S. Elwyn, and N. A. Moran. 1999. Calibrating bacterial evolution. Proc. Natl. Acad. Sci. USA **96**: 12638–12643.

Oleksiak, M. F., G. A. Churchill, and D. L. Crawford. 2002. Variation in gene expression within and among natural populations. Nat. Genet. **32**:261–266.

Papp, B., C. Pal, and L. D. Hurst. 2003. Evolution of cis-regulatory elements in duplicated genes of yeast. Trends Genet. **19**:417–422.

Pratt, J. M., J. Petty, I. Riba-Garcia, D. H. L. Robertson, S. J. Gaskell, S. G. Oliver, and R. J. Beynon. 2002. Dynamics of protein turnover, a missing dimension in proteomics. Mol. Cell. Proteomics **1**:579–591.

Rubin, G. M., M. D. Yandell, J. R. Wortman et al. (52 co-authors). 2000. Comparative genomics of the eukaryotes. Science **287**:2204–2215.

Segre, D., D. Vitkup, and G. Church. 2002. Analysis of optimality in natural and perturbed metabolic networks. Proc. Natl. Acad. Sci. USA **99**:15112–15117.

Sharp, P. M., and W.-H. Li. 1986. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15**:1281–1295.

Sherman, F. 1991. Getting started with yeast. Methods. Enzymol. **194**:3–21.

Stryer, L. 1995. Biochemistry. Freeman, New York.

Townsend, J. P., D. Cavalieri, and D. L. Hartl. 2003. Population genetic variation in genome-wide gene expression. Mol. Biol. Evol. **20**:955–963.

Varshavsky, A. 1996. The N-end rule: functions, mysteries, uses. Proc. Natl. Acad. Sci. USA **93**:12142–12149.

Wang, Y. L., C. L. Liu, J. D. Storey, R. J. Tibshirani, D. Herschlag, and P. O. Brown. 2002. Precision and functional specificity in mRNA decay. Proc. Natl. Acad. Sci. USA **99**:5860–5865.

Wittkopp, P. J., B. K. Haerum, and A. G. Clark. 2004. Evolutionary changes in cis and trans gene regulation. Nature **430**:85–88.

Kenneth Wolfe, Associate Editor

Accepted February 28, 2005