Expression Level, Evolutionary Rate, and the Cost of Expression

Joshua L. Cherry*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

*Corresponding author: E-mail: jcherry@ncbi.nlm.nih.gov.

Accepted: 27 September 2010

Abstract

There is great variation in the rates of sequence evolution among proteins encoded by the same genome. The strongest correlate of evolutionary rate is expression level: highly expressed proteins tend to evolve slowly. This observation has led to the proposal that a major determinant of protein evolutionary rate involves the toxic effects of protein that misfolds due to transcriptional and translational errors (the mistranslation-induced misfolding [MIM] hypothesis). Here, I present a model that explains the correlation of evolutionary rate and expression level by selection for function. The basis of this model is that selection keeps expression levels near optima that reflect a trade-off between beneficial effects of the protein's function and some nonspecific cost of expression (e.g., the biochemical cost of synthesizing protein). Simulations confirm the predictions of the model. Like the MIM hypothesis, this model predicts several other relationships that are observed empirically. Although the model is based on selection for protein function, it is consistent with findings that a protein's rate of evolution is at most weakly correlated with its importance for fitness as measured by gene knockout experiments.

Key words: expression level, protein evolution, population genetics, molecular evolution.

Introduction

It has long been known that there is considerable variation in the rate of sequence evolution among proteins encoded by a genome (Kimura 1986). This variation is much larger than the variation in synonymous substitution rates, so an explanation involving variation in mutation rates seems unlikely. These observations have been abundantly confirmed with the availability of more sequence data, including sequences of entire genomes (see, e.g., Makalowski and Boguski 1998; Waterston et al. 2002; Stein et al. 2003). An explanation of protein evolution should account for within-genome differences, and the nature of these differences may help us to distinguish among such explanations.

Protein evolutionary rate is correlated with many other variables, many of which also correlate with each other. The causal connections among these variables are unclear. However, an important fact has emerged: the best predictor of a protein's evolutionary rate is its expression level (Pál et al. 2001; Krylov et al. 2003; Rocha and Danchin 2004; Drummond et al. 2006; Drummond and Wilke 2008). Specifically, more highly expressed proteins tend to have lower evolutionary rates. Measures of a protein's contributions to fitness, such as the apparent cost of gene disruption or the propensity of a gene to be lost over evolutionary time, are comparatively poor predictors of the rate of sequence evolution. This has led some to question the role of selection for function as a determinant of protein evolutionary rate and as a major constraint on protein evolution. It has also led to alternative hypotheses, most notably the suggestion that selection against the harmful effects of mistranslation-induced misfolding of proteins is the major determinant of evolutionary rate (the MIM hypothesis) (Drummond et al. 2005; Drummond and Wilke 2008).

Here, I present a model that accounts for the correlation between expression level and evolutionary rate in terms of selection for protein function. This model is similar to that recently proposed by Gout et al. (2010). The assumptions involve a cost of gene expression and diminishing returns for the production of any particular protein. If expression levels are optimal, the fitness cost of the loss of a small fraction of protein function will be approximately proportional to the protein's expression level. Thus, more highly expressed proteins will be subject to stronger selection for function, leading to greater constraints on protein sequence and a lower rate of protein sequence evolution.

Materials and Methods

Concrete Model

A genotype consisted of two or three parts, depending on whether codon choice was included in the model. A "protein sequence" was modeled by 1,000 bits that determined specific activity. An additional 12 bits determined expression level. For modeling selection on codon choice, an additional 1,000 bits were included, each associated with a particular "nonsynonymous" bit. For modeling selection for translational efficiency, it was only necessary to keep track of the total number of optimal codons because the positions of these codons were irrelevant to the overall translational efficiency.

The protein's specific activity was determined by the sequence as follows. Each of the 1,000 bits was assumed to make a certain contribution to the logarithm of specific activity. A number representing this contribution was assigned to each position. For most of the simulations, the ith bit was assigned a contribution g_i equal to $i/1{,}000$; the results presented in figure 6 involve different assumptions, which are specified in the corresponding text. The specific activity of the most active sequence, consisting of all ones, was taken to equal one. Each zero-valued bit decreased the specific activity by a factor of $\exp(g_i)$. Thus, if Z is the set of all indices for which the corresponding bit is zero, the specific activity is given by

$$\exp(-\sum_{i \in Z} g_i).$$

Expression level was taken to equal $\exp(0.005n)/10^9$, where n is an integer between 0 and 4,095 (inclusive) that is determined by the 12 bits that specify expression level. These 12 bits were interpreted as a Gray code for the integer. Specifically, the value 0 was represented by 12 zeros, and the representation of $n + 1$ was obtained from the representation of n by inversion of the lowest-order bit that yielded an encoding not already assigned to a smaller integer. The details of the encoding of expression level are not relevant to the theoretical predictions, which simply assume that expression level is optimized.

Fitness was given by equation 1. The cost factor c was taken to be 1, except where translational efficiency was assumed to depend on codon choice, in which case it depended on F_{op} according to equation 5.

The rates of mutation were taken to be the same for all bit positions in the genotype and to be independent of the genotype.

Simulations

Simulations were performed using the Python programming language along with the NumPy package (Oliphant 2007). Each simulation consisted of a series of steps, each corresponding to the fixation of a mutant allele, and thus the inversion of one bit of the genotype. In each step, all possible single-bit changes were evaluated for their fitness effects. A pseudorandom choice among the bits was made, with the probability of each possible change proportional to its fixation probability, calculated from equation 4. At each stage, the total rate of protein sequence change (proportional to the sum of fixation probabilities across positions) was recorded, along with the rate of change at the bits encoding expression level. The latter was important for calculating the mean lifetime of a state. The mean rate was computed as the average of the rate over all iterations, weighted by the mean lifetime of the state. Where relevant, the rate of synonymous change and the fraction of optimal codons were also recorded at each step. The first 1,000 steps of a simulation were not used for calculations of means. Simulations were run for 10,000–3,000,000 steps, depending on the model parameters.

Numerical Predictions

The decreasing curves in figure 2*a* represent the optimal level of expression as a function of specific activity. For any specific activity σ, the expression level giving the highest fitness was found by solution of $\frac{\partial w}{\partial \epsilon}=0$, with w given by equation 1. This equation was solved numerically with the "roots" function of the NumPy package. For each payoff function considered, solution for a range of values of specific activity yielded the corresponding curve in figure 2*a*.

All the other predictions presented are independent of the payoff function. Where relevant, predictions are calculated from a linearization of the payoff function under the assumption that the ratio of change in fitness to fractional change in specific activity is proportional to expression level (eq. 3). This assumption will hold approximately, according to the theoretical results presented here, when expression level is optimal. Each bit position is analyzed separately under this assumption. At any time, either a "1" or a "0" will be fixed at the focal position. The ratio of the time spent with a 1 fixed to the time spent with a 0 fixed is equal to $\exp(2Ns)$, with s reflecting the fitness effect of changing a 0 to a 1 (this follows from eq. 2 and the symmetry of mutation rates). Application of this approximation to each bit allows prediction of specific activity as a function of expression level (the rising curve in fig. 2*a*). Furthermore, the mean rate of evolution at any sequence position is proportional to

$$2\frac{1}{1+\exp(2Ns)}p_{\text{fix}}(s,N).$$

Summation of this quantity over all relevant positions (synonymous or nonsynonymous) yields an approximate prediction of the corresponding evolutionary rate as a function of expression level.

Results

The Model

The rate of a protein's evolution relative to the mutation rate depends on the distribution of fitness effect among mutant sequences. The rate is determined primarily by the distribution among mutants of small effect; mutants with large deleterious effects are extremely unlikely to fix, and there can be few reversals of such fixations if they rarely happen in the first place.

Consider, then, the fitness effect of a change to a protein sequence that alters its functionality by a small amount. Suppose, for specificity, that it decreases some measure of protein function by 1%. It might, for example, reduce an enzyme's activity by 1% due to a subtle alteration of the active site or cause 1% of the protein to fail to fold correctly. What will be the fitness effect of such a change? How might it relate to the protein's expression level?

For many proteins, a 1% decrease in activity due to an altered protein sequence will have roughly the same effect as the disappearance of 1% of the protein without any compensatory advantage. For a more highly expressed gene, this 1% corresponds to a proportionately larger quantity of protein. If the protein that is notionally lost has the same value per quantity for both genes, the total fitness loss will be proportionately higher for the more highly expressed gene. This equality of value of protein will hold approximately for mutations of small effect if expression levels are optimized by selection, as explained below.

The overall importance of a gene to fitness is not closely linked to its expression level. Genes with low expression levels may be essential and highly expressed genes might make only small contributions to fitness. A gene's total contribution to fitness and hence the total value of its product might be unconnected to expression level. The average value of a gene product—the ratio of its total fitness contribution to its expression level—would then decrease with expression level.

However, neither the total nor the average value of a gene product is directly related to the question at hand. What matters is the sensitivity of fitness to small changes in the amount of protein. In other words, what matters, in the limit of small effect size, is the marginal value of the gene product. Suppose that there is a general (non-gene–specific) cost to increasing expression, such as the metabolic cost of protein, as appears to be the case for *Escherichia coli* (Stoebel et al. 2008). When expression levels are optimal, the marginal values of different proteins will be identical. If this were not so, shifting protein production from products with low marginal value to those with high marginal value would be advantageous, so this would not be an optimum after all. Selection would be expected to keep expression levels near their optimal values, and adaptation of this sort has been demonstrated in laboratory experiments (Dekel and Alon 2005). Thus, for changes to the protein sequence that have a small effect on function, the effect on fitness should be approximately proportional to the expression level.

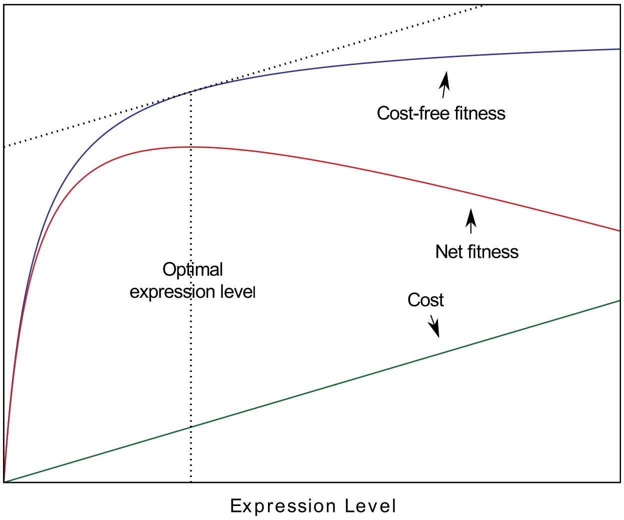

Fig. 1.—The relationship between expression level and fitness. A cost-free fitness function with diminishing returns is shown in blue. The total fitness cost of protein is proportional to the expression level, as shown in green. Fitness, equal to the difference between these, is shown in red. The peak of the red fitness curve occurs where the slope of the blue curve equals the slope of the green line. Thus, the tangent to the blue curve is necessarily parallel to the green line. This relationship holds regardless of the shape of the blue curve.

Essentially the same argument can be given from the point of view of a single gene whose expression comes at a cost that is unrelated to the protein's function. This argument is given below. The formal model that emerges is then used as a basis for computer simulations and numerical analyses that confirm the argument given above and demonstrate other predictions about the evolution of coding sequences.

Suppose that the contribution of protein function to fitness increases with expression level but with diminishing returns (the blue curve in fig. 1). A relationship of this type may be derived, for example, from theoretical results concerning fluxes through metabolic pathways (Kacser and Burns 1973; Hartl et al. 1985). Suppose also that expression comes with a proportional cost (the green line in fig. 1). There will then be an evolutionary trade-off between the benefit of protein function and the cost of expression (the red curve). At low expression levels, increasing expression will have a net beneficial effect because the benefit of increased protein function will outweigh the cost of the additional expression. At high expression levels, it will be favorable to decrease expression because the decrease in the corresponding cost will outweigh the lost functional benefit. The optimum expression level will be intermediate, at the point where the marginal benefit of protein function equals the cost of additional expression (the vertical line in fig. 1). Consider a gene expressed at this optimal expression level, ϵ_{opt}. Decreasing the expression level by an infinitesimal fraction δ (changing it to $(1-\delta)\ \epsilon_{opt}$,) will leave fitness unchanged; there will be a fitness loss due to loss of protein function, equal to $c\epsilon_{opt}\ \delta$ for some constant c, but a compensatory fitness gain of $c\epsilon_{opt}\ \delta$ due to the decreased cost of expression. Now consider a change to the protein sequence that causes the loss of a fraction δ of protein function (e.g., decreases the specific activity of an enzyme). This will incur the same fitness loss of $c\epsilon_{opt}\ \delta$ due to the loss of protein function but without the compensatory decrease in the cost of expression. Thus, the net effect will be to decrease fitness by a quantity that is proportional to the expression level and, therefore, larger for more highly expressed genes. Similarly, the fitness gain due to a fractional increase in protein function will be larger for more highly expressed genes. Therefore, the fitness effects of changes to the protein sequence are expected to be larger for more highly expressed genes, leading to lower rates of protein sequence evolution.

More formally, assume that, neglecting the cost of expression, fitness is given by a function $f(\alpha)$, where α is a quantity that I will call "activity." The (total) activity is proportional to both the expression level ϵ and the "specific activity" σ, which is determined by the protein sequence. Specifically, $\alpha = \sigma\,\epsilon$. I will refer to the function f as the "payoff function."

Note that activity is used here in a broader sense than its strict enzymological meaning. It is meant to be applicable even to proteins that are not enzymes. Furthermore, even for an enzyme it need not correspond to what a biochemist would call "enzyme activity," which is proportional to k_{cat}. For example, on the basis of metabolic control analysis (Kacser and Burns 1973), fitness has been modeled as a function of a quantity that is proportional to k_{cat}/K_M (Hartl et al. 1985). Also, although the parameter σ is referred to as specific activity, it can reflect such factors as the failure of a fraction of the protein to fold correctly.

The cost of protein production is taken to be proportional to the expression level ϵ. Fitness is then given by

$$w(\sigma,\varepsilon) = f(\sigma\varepsilon) - c\varepsilon \tag{1}$$

Because the scale on which we measure expression level is arbitrary, it will sometimes be convenient to take the cost factor c to equal one. To model the effect of codon choice on the cost of expression, we can allow c to depend on codon usage.

Fitness might instead be given by an expression such as $f(\sigma\epsilon) \times (1 - c\epsilon)$ or $f(\sigma\epsilon) \times \exp(-c\epsilon)$. On a related note, we should distinguish between the ordinary fitness and its logarithm, the Malthusian fitness. Such details have negligible effects on the numerical results presented below, for which $c\epsilon << 1$ (as expected for all but very highly expressed proteins) and $w(\sigma,\varepsilon) \approx 1$. More importantly, these details are not relevant to the general argument presented above.

One might also use a nonlinear cost function such as that proposed by Dekel and Alon (2005) based on experiment. However, use of this function has negligible effect on the numerical results presented below. This is so because the deviations from linearity only become significant at extremely high expression levels and are negligible in the relevant range of expression. For the vast majority of real genes, expression levels will also be too low for such deviations to be significant. Moreover, a nonlinear cost function is likely inappropriate for gene-wise treatment of genes in the same genome. This is because, on the simplest model, high expression of a gene will increase the cost of expression of all genes, leaving the relative marginal costs unchanged. This expectation is largely borne out by the results of Stoebel et al. (2008).

If fitness is given by equation 1, the partial derivatives of fitness are given by

$$\frac{\partial w}{\partial \varepsilon}(\sigma,\varepsilon) = \sigma f'(\sigma\varepsilon) - c \tag{2a}$$

and

$$\frac{\partial w}{\partial \sigma}(\sigma,\varepsilon) = \varepsilon f'(\sigma\varepsilon). \tag{2b}$$

From equation 2a, it follows that when expression level is optimal, $f'(\sigma\epsilon)=c/\sigma$. From equation 2b, it then follows that $\frac{\partial w}{\partial \sigma}(\sigma,\epsilon)=c\epsilon/\sigma$ when expression level is optimal. Equivalently, the fitness effect of a small change in specific activity $\Delta\sigma$ is given approximately by

$$\Delta w \approx c\varepsilon\frac{\Delta\sigma}{\sigma} \tag{3}$$

Thus, the sensitivity of fitness to a small change in specific activity is proportional to the gene's expression level.

The above considered a fixed specific activity. In reality, the specific activity changes as the protein sequence evolves. Suppose that the expression level were held constant. Fitness would then be an increasing function of specific activity that is similar to the blue curve in figure 1. The higher the specific activity, the greater the tendency of fixations to be deleterious. There are two reasons for this. First, as the slope of the curve decreases, selection for activity becomes weaker (Hartl et al. 1985; Cherry 1998). Second, as the protein sequence becomes more adapted, mutation is increasingly biased toward maladaptive changes (this is true, at least, on many reasonable models). Specific activity will tend to evolve toward a value at which advantageous and deleterious fixations are balanced (Cherry 1998).

The level of expression and the protein sequence evolve in concert, each affecting the selective force acting on the

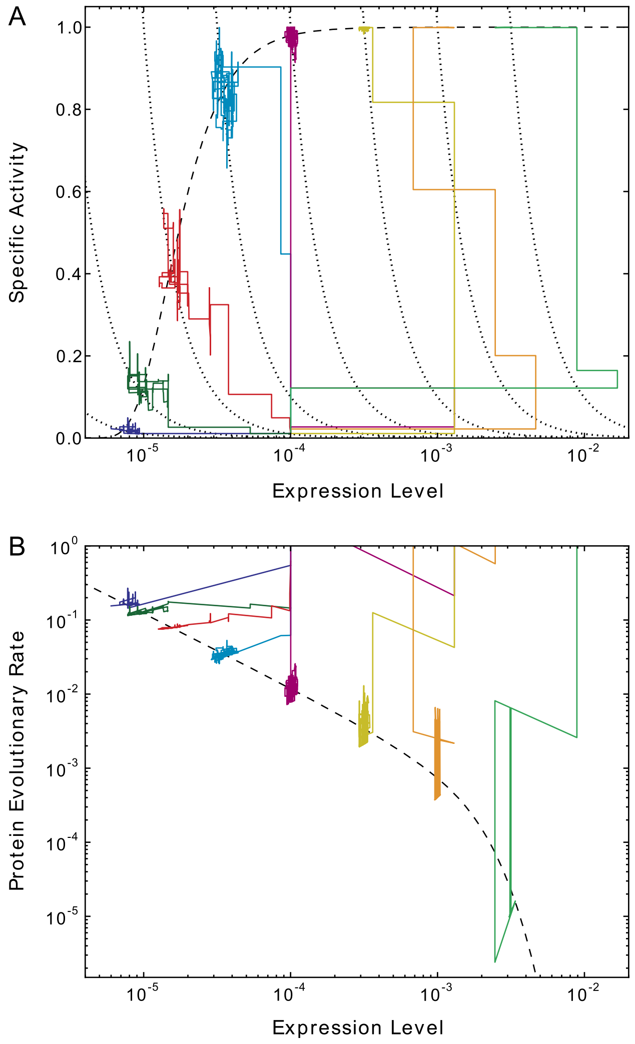

Fig. 2.—Evolution in simulation results. The trajectories of simulation results are shown for payoff functions of the form $\alpha/(\alpha + K)$ for $K = 10^{-12}, 10^{-11}, \ldots 10^{-5}$. The plots show the first 500 steps of each simulation. Each step corresponds to a single sequence change, affecting either the protein sequence or the expression level. (*A*) Expression level and specific activity. Each sequence change produces either horizontal or vertical movement, corresponding to a change in expression level or protein sequence, respectively. Each trajectory is drawn toward the intersection of the corresponding falling curve (specific to the benefit function) and the single rising curve. (*B*) Expression level and protein evolutionary rate. Changes to the protein sequence again yield vertical movement because they do not affect expression level. Changes to the expression level yield diagonal movement because they also affect the rate of protein evolution. The curve represents a numerical prediction that is independent of the payoff function.

other. Evolution will tend to keep the expression level and specific activity in a certain region where two conditions hold. First, the expression level is approximately optimal for the specific activity. Second, the specific activity has roughly the value it would tend to evolve toward if expression level was fixed at that value. These two conditions are expected to hold under fairly general conditions. When fitness takes the form of equation 1, the protein sequence will behave as though equation 3 applied. The specific activity will therefore tend to increase with expression level, and the rate of protein evolution will decrease.

Numerical Confirmation

In order to explore concrete instances of this model, I consider a simplified model of a protein sequence. A protein sequence is modeled as a sequence of 1,000 bits. The highest possible specific activity is 1, which is achieved by a sequence of all "ones." Every "zero" diminishes the specific activity by a factor that depends on its position in the sequence. Equivalently, every sequence position is assigned a value that specifies the logarithm of its multiplicative effect on specific activity. I initially assume that the natural logarithm of the contribution of the *i*th bit is given by *i*/1,000 ($i = 1, 2 \ldots 1{,}000$). Other possibilities are explored below.

The expression level is also assumed to be encoded by the genotype. It is determined by a sequence of 12 bits, which are interpreted as a Gray code representation of an integer between 0 and 4,095. The use of the Gray code guarantees that an increment or decrement by one unit can always be achieved by the change of a single bit (a single "mutation"), while allowing other bits to have larger effects. The expression level is an exponential function of the integer *n*, namely $\exp(0.005n)/10^9$. Thus, changing the integer by 1 changes the expression level by approximately 0.5%.

The population was assumed to be a haploid Wright–Fisher population. It was assumed that each newly arising allele goes to fixation or extinction before new mutant alleles enter the population (the "weak mutation" approximation). Therefore, the fixation probability of a newly arising allele is given approximately by

$$p_{\text{fix}}(s, N) = \frac{1 - \exp(-2s)}{1 - \exp(-2Ns)}, \qquad (4)$$

where *N* is the population size and *s* is the selection coefficient (Kimura 1957). The population size was taken to be 10^6. The selection coefficient was calculated as the natural logarithm of the ratio of the fitness of the novel genotype to that of the established allele.

Simulations consisted of a series of steps in which an established genotype was replaced by one of its single-bit variants. Each variant's probability of being chosen as the replacement was proportional to its fixation probably as given by equation 4, with selection coefficients calculated from equation 1. These equations also formed the basis of approximate theoretical predictions, which can be compared with the simulation results.

Figure 2*a* shows the evolution of expression level and specific activity for payoff functions of the form $\alpha/(K + \alpha)$. From a common starting point, each simulation moves fairly rapidly to a different region of the plane, where it largely remains. The rising curve gives a prediction of the specific activity to which the system will evolve for a fixed expression level. It is based on a linearization that assumes that the selective effect of a change in specific activity is proportional to the level of expression (eq. 3), which is predicted to hold approximately by the argument given above. Each falling curve is specific to a payoff function. It gives the expression level that maximizes fitness, according to equation 1, for any specific activity. Its intersection with the rising curve represents the simultaneous satisfaction of two conditions that are expected to hold approximately at evolutionary steady state. For each falling curve, the corresponding simulation evolves to the vicinity of this point, as predicted by theory.

Figure 2*b* shows, for the same simulations, the evolution of expression level and protein evolutionary rate. Each simulation again evolves toward a different region of the plane. These regions lie along a falling curve that represents a theoretical approximation, based on equation 3, that is independent of the payoff function. Payoff functions that result in higher expression levels also result in lower evolutionary rates. Thus, this figure illustrates the fundamental prediction of the argument: protein evolutionary rate decreases as expression level increases.

In all the simulations whose results are show in figure 2, the payoff function had the form $\alpha/(K+\alpha)$. Thus, the payoff functions had the same shape, differing only by scaling. In particular, the gene being modeled was in all cases essential to fitness: the absence of protein activity would lead to zero fitness, even neglecting any cost of the protein. Theory predicts that the rate of a protein's evolution should be largely unaffected by the payoff function, except through the effect of the payoff function on the optimized expression level. If their expression levels are the same, a protein essential to fitness should have approximately the same evolutionary rate as a protein that makes only a small contribution to fitness. This prediction can be confirmed by simulations using a wider variety of payoff functions.

Representatives of four families of payoff functions are plotted in figure 3. In three of the families, the contribution to fitness has a Michaelis–Menten form as above, but the total possible contribution to fitness varies. These payoff functions have the form

$$f(\alpha) = (1-d) + d\frac{\alpha}{K+\alpha}. \qquad (5)$$

The parameter *d* dictates the maximum contribution to fitness that can be made by a gene and puts an upper bound on the fitness cost of deletion of the gene. Families with *d* equal to 1, 0.1, and 0.01 are represented in figure 3 (blue, yellow, and red curves, respectively) and are used in the simulations discussed next. In addition, a family with an exponential approach to maximum payoff is considered:

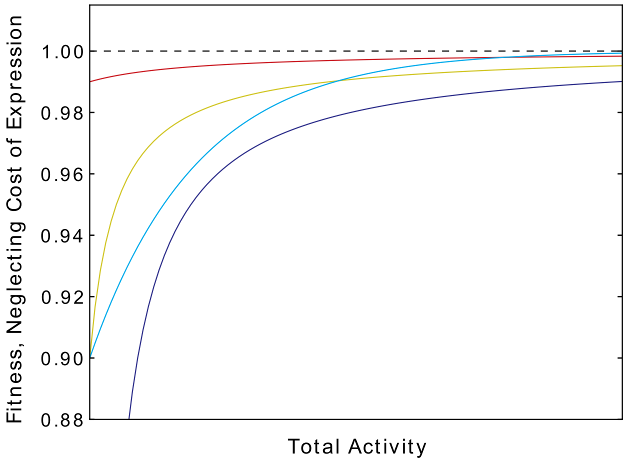

Fig. 3.—Representatives of four families of payoff functions. Fitness, neglecting the cost of protein, is plotted as a function of total activity for four different payoff functions. Each plotted function is a representative of a family related by scaling in the horizontal dimension. These families are specified in the text.

$$f(\alpha) = (1-d) + d(1-\exp(-\alpha/K)), \qquad (6)$$

with $d = 0.1$ (cyan curve in fig. 3).

Figure 4 shows results for simulations for all four families of payoff functions. Results roughly fall along the same theoretical curve, which relates evolutionary rate to expression level. To a good approximation, the evolutionary rate depends only on the expression level, not on the payoff function. Thus, a highly expressed gene making only a small contribution to fitness has a lower evolutionary rate than an essential gene with a lower expression level. Consider, for example, the gene represented by the rightmost red square in figure 4. Disruption of this gene would lead to less than a 1% loss of fitness. Nonetheless, the protein sequence is under stronger selection and evolves more slowly than most of the others represented in the figure, including several that are essential to fitness. Conversely, some essential genes (blue circles) are among the most rapidly evolving genes.

The Effect of the Distribution of Mutational Effects

The downward trend illustrated by figure 4 is expected under fairly general conditions. The particular form of the relationship, however, depends on how protein sequence maps to specific activity. Over part of its range, the curve in figure 4 approximates a straight line with slope equal to −1, corresponding to an evolutionary rate that is approximately proportional to the reciprocal of expression level. As expression level becomes large, evolutionary rate begins to drop precipitously. This behavior can be understood in terms of the model for protein function. Any position in the sequence has an associated selection coefficient that,

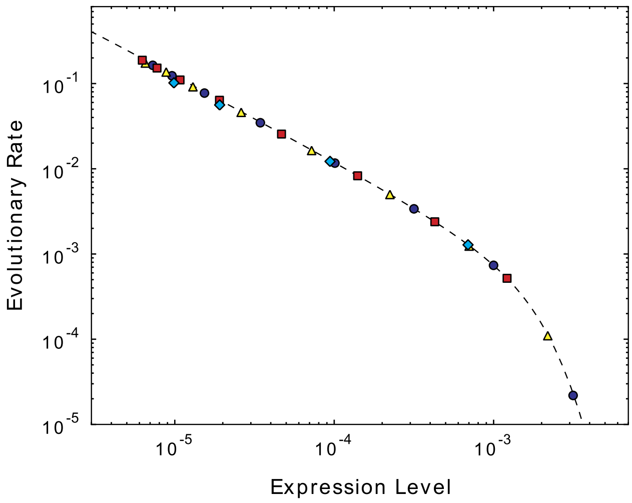

Fig. 4.—Evolutionary rate as a function of expression level for a wide variety of payoff functions. Expression levels and evolutionary rates observed in simulations are plotted, along with a theoretical curve. Each point represents the results of a simulation using a particular payoff function. These functions come from the four families illustrated in figure 3, with plot symbols colored accordingly. Blue circles: equation 5 with $d = 1$ and $K = 10^{-12}, 10^{-11}, \ldots 10^{-5}$. Yellow triangles: equation 5 with $d = 0.1$ and $K = 5 \times 10^{-12}, 5 \times 10^{-11}, \ldots 5 \times 10^{-5}$. Red squares: equation 5 with d = 0.01 and K = $2 \times 10^{-11}, 2 \times 10^{-10}, \ldots 2 \times 10^{-4}$. Cyan diamonds: equation 6 with $d = 0.1$ and $K = 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}$.

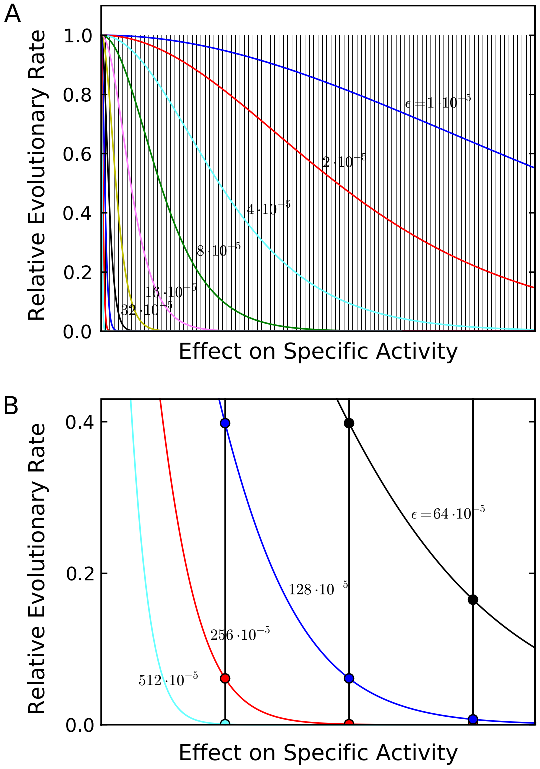

Fig. 5.—Basis for the shape of the relationship between expression level and evolutionary rate. Each curve corresponds to a particular expression level and gives the relative evolutionary rate as a function of effect size. Moving right to left, expression level increases by a factor of two with each curve. (*A*) At intermediate expression levels, the total evolutionary rate is approximately proportional to the area under the curve. Doubling the expression level approximately halves the evolutionary rate. (*B*) At high expression levels (narrow curves), the total evolutionary rate falls more rapidly as expression level increases due to the discrete nature of the distribution of effect sizes.

according to the model, is proportional to the level of expression. The rate of evolution at that site is a decreasing function of the selection coefficient. It was assumed above that the effect of position *i* on the logarithm of specific activity is proportional to *i*. Thus, the distribution of effect sizes among sites can be approximated by a uniform distribution for sufficiently small effect size, as illustrated in figure 5. Also shown in the figure are curves relating the evolutionary rate at a site to its effect on specific activity. Each curve corresponds to a different expression level: increasing expression level contracts the curve horizontally. Over a wide range of expression levels, the overall evolutionary rate is approximately proportional to the area under the corresponding curve. Evolutionary rate is therefore proportional to the reciprocal of expression level. For very high expression levels (corresponding to narrow curves), the continuous approximation breaks down and the evolutionary rate falls rapidly as expression level increases.

Figure 6 shows theoretical predictions and simulation results for four assumptions about the effects of different sites on specific activity:

- Blue circles: As in the simulations above, the effect of the *i*th site on the logarithm of specific activity is equal to *i*/1,000.
- Green squares: The effect of the *i*th site is equal to *i*/10,000. Because the effect sizes are all 10-fold smaller, the curve is similar but is shifted to the right.
- Red triangles: The effect of the *i*th site is proportional to i^2. Reasoning analogous to that illustrated in figure 5 may be applied to this case. The curve is again predicted to be well approximated by a straight line over a range of expression levels. However, due to the i^2 dependence, the slope is predicted to be −1/2. Figure 6 bears out this prediction.
- Cyan diamonds: The effect of the *i*th site is proportional to $i^{1/2}$. In this case, a slope of −2 is expected for the approximately linear portion of the curve. This is also confirmed by figure 6.

Despite very different assumptions about the effects of mutations on specific activity, the four cases yield the same qualitative result: evolutionary rate decreases with expression level, as predicted by the general model. Furthermore, even when all of these results are combined, there is a strong negative correlation between expression level and evolutionary rate.

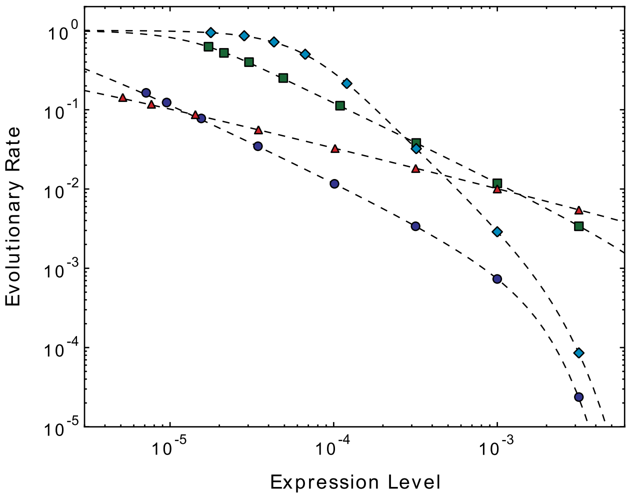

FIG. 6.—Evolutionary rate and expression level for different distributions of effect size. Predicted and observed rates are shown for four assumptions about the distribution of effect sizes among sequence positions. The distributions are specified in the text.

Synonymous Codons

The analysis presented above does not consider selection among synonymous codons. There are two reasons for considering this type of selection. First, it might affect the relationship between expression level and the rate of nonsynonymous change. This effect turns out to be small under the models of synonymous selection explored here. Second, predictions may be made about additional relationships, such as the relationship between synonymous and nonsynonymous rate. The model analyzed here predicts several such relationships that are observed empirically and predicted by a model of the MIM hypothesis.

In order to model a degenerate genetic code, I assume that each of the 1,000 bits in the model protein sequence is paired with an additional bit that is analogous to a fully degenerate site in a coding sequence. In essence, a codon is modeled as a pair of bit positions: a nonsynonymous position and a synonymous position. The synonymous position affects either the efficiency of translation (and hence the cost of expression) or translational accuracy (and hence the mean specific activity of the protein produced). A 1 at the synonymous position corresponds to a preferred (more efficiently or more accurately translated) codon and a 0 to a nonpreferred codon.

To model selection for translational efficiency, I assume that the cost factor c of equation 1 depends on the fraction of optimal codons, F_{op}, as follows:

$$c = F_{\text{op}}k_1 + (1 - F_{\text{op}})k_0, \qquad (7)$$

with $k_1 = 2/3$ and $k_0 = 4/3$. With these parameters, $c = 1$ for an equal mix of preferred and nonpreferred codons ($F_{\text{op}} = 1/2$), and the cost of expression with $F_{\text{op}} = 1$ (all preferred codons) is half the cost of expression with $F_{\text{op}} = 0$ (all nonpreferred codons).

Figure 7*a* shows numerical predictions and simulation results for this model of selection for translational efficiency. As the figure shows, the protein's evolutionary rate (the nonsynonymous rate) again decreases with expression level. The form of the relationship is slightly different. This reflects the fact that the total cost of a protein increases less than linearly with its expression level because more highly expressed genes have more nearly optimal codon usage at equilibrium. Figure 7*a* also shows that, as expected, the rate of synonymous evolution decreases with expression level. The decrease in synonymous rate is fairly abrupt on the logarithmic scale. This reflects the fact that the magnitude of the selection coefficient is identical for all synonymous changes. On a more realistic model, different synonymous changes would have different effects on fitness and the decrease in synonymous rate with expression would be somewhat more gradual.

Figure 7*b* shows results under the assumption that the synonymous site affects translational accuracy. Specifically, it was assumed that a 0 at the synonymous site leads to a 1% translational error at the corresponding amino acid position, whereas a 1 yields error-free translation at that position. Again both the nonsynonymous and synonymous rates decrease with expression level. The relationship between nonsynonymous rate and expression level is largely unchanged by the introduction of synonymous sites and translational error. In fact the predicted relationship shown in figure 7*b* was calculated without consideration of translational error. The decline in synonymous rate with expression level is more gradual than in figure 7*a* because there is great variation in selection coefficient among synonymous sites: those associated with more important amino acid positions are more strongly selected for translational accuracy.

Figure 8 shows the pairwise relationships among several variables for both models of synonymous selection. Increasing relationships are colored yellow and decreasing relationships are colored cyan. These relationships may be compared with those in Drummond and Wilke (2008). If d*S*/d*N* is substituted for the transition:transversion ratio (see Discussion), the signs of all ten pairwise relationships are reproduced with either model of synonymous selection.

Discussion

The model presented here, which is similar to that proposed by Gout et al. (2010), explains the negative correlation between expression level and protein evolutionary rate in terms of selection for protein function. With the addition of either of two models for selection on codon usage, several other relationships that have been empirically observed can be produced by the model.

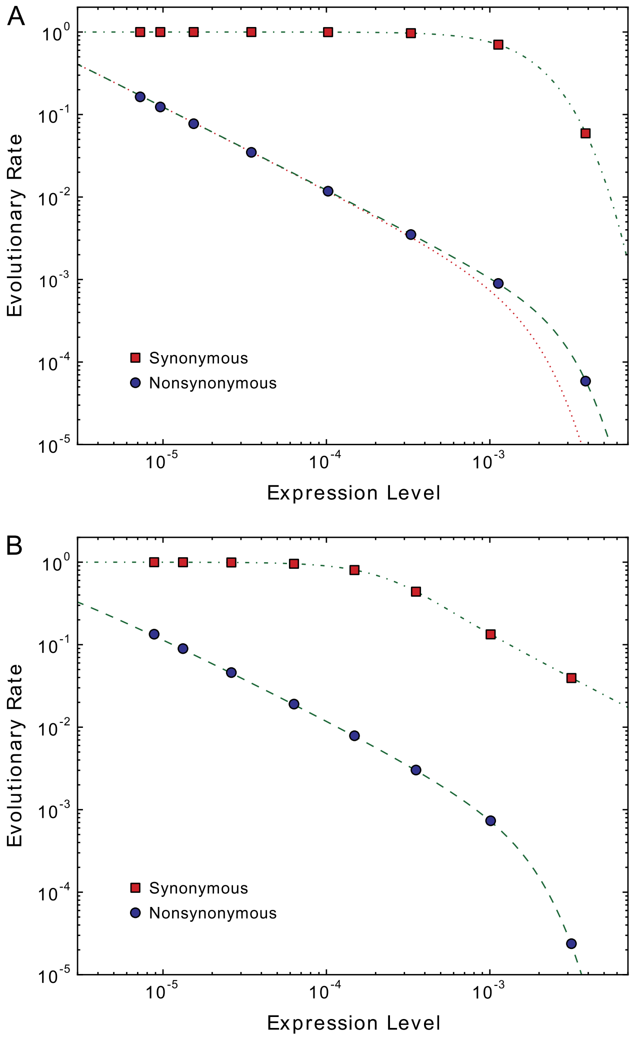

Fig. 7.—Synonymous and nonsynonymous rates. Predicted and observed relationships between expression level and synonymous and nonsynonymous evolutionary rates are shown for two assumptions about selection for codon usage. (A) Selection for translational efficiency. (B) Selection for translational accuracy.

The model assumes that there is a cost to expression and that expression levels are approximately optimized by natural selection. Under these conditions, the marginal values of all gene products will be identical, even if their total contributions to fitness are very different. This principle of optimization is familiar in other contexts, such as ecology and economics. One economics textbook (Gwartney et al. 2008, p. 422) sums it up as follows: "the consumer will maximize his or her satisfaction (or total utility) by ensuring that the last dollar spent on each commodity yields an equal degree of marginal utility." As with commodities, so with gene products. Loosely speaking, fitness is maximized when the organism allocates costly expression such that the last molecule "spent" on each protein yields an equal fitness benefit. A change to a protein's sequence that decreases its functionality by a small fraction will, by assumption, have the same effect on fitness as the loss of the same fraction of the protein. For a more highly expressed gene, this corresponds to the loss of a larger number of approximately equally valuable protein molecules. Thus, selective constraints on protein sequences are stronger for more highly expressed proteins.

Although cost plays a central role in the model, the fitness effect of a change to the protein sequence derives solely from the resulting gain or loss of protein function, not from a change in the total cost of expression. The sequence change does not itself alter the expression level or the associated cost, even though it may create selective pressure for subsequent changes in expression level.

The correlation does not result from a simple direct effect of expression level on evolutionary rate. If the expression level of a gene were somehow held artificially high throughout evolution, the consequence would not be a correspondingly lower rate of protein evolution. In fact the opposite would be true: the evolutionary rate would increase due to relaxed selection on the more abundant protein. The predicted relationship holds only when expression level is optimized.

Although it is based on selection for function, the model does not imply a relationship between a protein's evolutionary rate and the contribution of the protein's function to fitness (fig. 4). It is therefore compatible with evidence that there is little relationship between a protein's evolutionary rate and its importance to fitness, except perhaps through their mutual correlation with expression level (Hurst and Smith 1999; Pál et al. 2003; Rocha and Danchin 2004; Drummond et al. 2006). This is so because the rate of sequence evolution under purifying selection depends mainly on the fitness effects of small fractional changes to the amount of function, and the effect of complete loss of a protein's function is not directly relevant to this (the marginal contribution to fitness is not determined by the total contribution to fitness). Thus, evidence that a gene's evolutionary rate is not related to its functional importance is no reason to abandon the hypothesis that selection for function is the main constraint on protein sequence evolution.

Assumptions and Variations

The predictions of the model rest on several assumptions. Although these assumptions plausibly hold for many genes, some genes undoubtedly violate them. Such violations deserve consideration, but they do not invalidate the model as an explanation of a general trend.

Certain assumptions were made about how a protein's sequence and its expression level together determine its functional contribution to fitness. It was assumed that, except for the cost of the additional protein, a genotype that

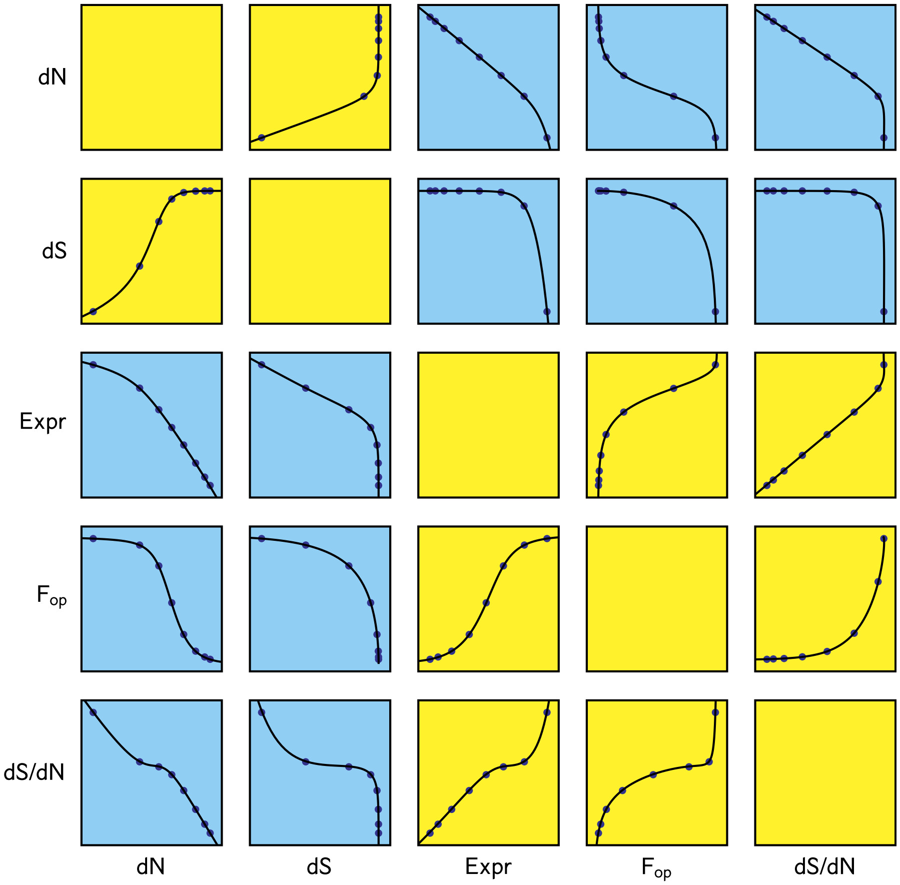

Fig. 8.—Predicted and observed relationships among several variables. Relationships between pairs of variables (d*N*, d*S*, expression level, F_{op}, and d*S*/d*N*) are shown for two models of selection for codon usage: a model of selection for translational efficiency (above and to the right of the main diagonal) and a model of selection for translational accuracy (below and to the left of the main diagonal). Yellow coloring indicates an increasing relationship, and cyan indicates a decreasing relationship.

produces a large amount of relatively inactive protein is equivalent to one that produces a smaller amount of correspondingly more active protein. This assumption is embodied in equation 1, where the specific activity appears only in its product with expression level. Where stoichiometry is critical, this assumption may not hold. This assumption is likely to be an excellent approximation for most enzymes, as suggested by the usual approximations of enzyme kinetics, according to which the reaction velocity depends on the specific activity and the quantity of enzyme only through their product. The model may apply even to cases where stoichiometry is important. Suppose that the main way in which changes to the protein sequence lead to loss of function is by causing some fraction of the protein to fail to fold properly. Equation 1 may then apply because the quantity of functional protein is proportional to the product of the fraction of proper folding and the expression level.

It was also assumed that, although there are diminishing returns for increases in a protein's activity, additional activity can only improve fitness (neglecting any cost of any additional protein). For many proteins, this will not be the case: increasing the total activity beyond some point may decrease fitness, even if it comes at no cost in terms of protein production. Still, even for such proteins the model may apply. Consider figure 1. It is necessary for the argument that the cost-free fitness function—the blue curve—is increasing in the vicinity of the optimal expression level. However, suppose that this curve turns downward at some point far to the right, as it might well do in the region to the right of the plot. This will have no effect on the location of the expression optimum or the slope of the curve at that point. Nor would it invalidate the linear approximation for small relative changes in activity. It therefore has no effect on the conclusion that the strength of selection is approximately proportional to the expression level for mutations with small effect. For some genes, however, selection may bring the total activity to the vicinity of a peak in the payoff function, so that the conclusions will not hold. Genes encoding regulatory proteins may fall into this category.

These considerations suggest that the model will not apply to every gene. However, it need only apply to a significant fraction of genes in order to explain the correlation between expression level and evolutionary rate, which is far from perfect.

For genes to which the model applies, the form of the relationship between expression level and evolutionary rate will depend on how protein sequence maps to specific activity. This mapping might be called the activity landscape (by analogy to the fitness landscape). The argument illustrated in figure 1 tells us that the fitness effect of a small change in specific activity depends, to a first approximation, only on the gene's expression level. For any particular activity landscape, the evolutionary rate will then be determined, approximately, by the expression level alone, as illustrated in figure 3. However, differences in the nature of the landscape could lead to different evolutionary rates for genes with the same expression level. As illustrated by figure 6, differences of this sort need not mean that there is no correlation between expression level and evolutionary rate. They would imply only that genes with the same level of expression can differ in their evolutionary rates. Some such variation among genes with similar expression levels is allowed—indeed required—by the imperfect correlation between expression level and evolutionary rate. Differences of this sort would account for a component of rate variation that is specific to ortholog pairs but unconnected with expression level. According to one analysis (Wolf et al. 2010), a substantial fraction of rate variation is of this type.

Nature of the Cost

An obvious cost of gene expression is the metabolic cost of the protein. Protein constitutes approximately half of the dry weight of bacteria, and polymerization of amino acids consumes approximately half of the ATP equivalents required for growth (Gottschalk 1986, p. 38–39; Neidhardt et al. 1990, p. 4). Furthermore, increasing the expression level of a protein consumes some of the protein synthetic capacity of the cell, decreasing the overall rate of protein synthesis or requiring increased production of the expensive translational apparatus. To the extent that codon bias is the result of selection for translational efficiency, it reflects this cost.

Other costs of expression might also be invoked, with similar consequences. These include harmful effects of the protein product, such as the toxicity of misfolded proteins invoked by the MIM hypothesis. If this cost does not vary as the protein sequence changes, the model presented here applies fully; selective constraints on the protein sequence reflect selection for protein function, and the role of toxicity is only to affect the optimal expression level. A more general model could allow changes to the protein sequence to affect both the protein's function and the cost of expression.

Because different proteins have different half-lives, proteins synthesized at the same rate may be present at different steady-state levels. We may ask which of the two quantities—the rate of a protein's synthesis or the protein's concentration—is predicted to be more closely related to its evolutionary rate. This question is particularly important because of an argument advanced by Drummond et al. (2005) concerning the correlation between expression level and evolutionary rate. They argue against explanations that are based on selection for protein function on the grounds that the rate of synthesis of a yeast protein (as inferred from mRNA abundance) is a better predictor of its evolutionary rate than is the protein's abundance. This argument rests on the assumption that any explanation based on selection for function would imply that protein concentration is the better predictor. The model presented here suggests otherwise. To the extent that the cost of expression is dominated by such factors as the energetic cost of amino acid polymerization and the use of the translational apparatus, as appears to be the case for *E. coli* (Stoebel et al. 2008), the model predicts that the rate of synthesis will be the better predictor of evolutionary rate. The consequences of the cost of the amino acids are somewhat more difficult to assess; they depend on how quickly the constituent amino acids of a short-lived protein are made available for reuse (they might, e.g., linger in the form of short peptides). Different nonmetabolic costs would make different predictions. In short, the model presented here cannot be rejected on the basis of the argument made by Drummond et al.

The cost of a given amount of expression may vary with cell type, developmental stage, and environmental conditions. The strength of selection on protein sequence implied by expression under these different conditions would vary accordingly. Suppose, for example, that the cost of expression is particularly high in neurons, whether due to a greater sensitivity to toxic effects of misfolded proteins or to some other factor. The model would then explain the observation (Drummond and Wilke 2008) that expression in neurons is more strongly associated with low evolutionary rate than is expression in other tissues.

Synonymous Codons

A simplified model of a degenerate genetic code allowed exploration of selection for codon choice. The effects of both selection for translational efficiency (Ikemura 1981) and selection for translational accuracy (Akashi 1994) were considered. As figure 7 demonstrates, the rate of synonymous substitution decreases as expression level increases with either type of selection. The reason for this behavior in the presence of selection for efficiency is familiar: a codon that causes translation to be more costly will have a greater fitness cost if it is translated more frequently (Sharp and Li 1986). With selection for accuracy, the synonymous rate decreases with expression level for the same reason that the nonsynonymous rate decreases: changes to the protein sequence tend to be more unfavorable for highly expressed proteins, whether they are due to nonsynonymous changes or to translational errors.

Synonymous selection has only a small effect on nonsynonymous selection. As figure 7*a* shows, selection for translational efficiency causes the nonsynonymous rate to fall off less rapidly at high expression levels. Codon usage tends to be more favorable at higher expression levels, so the total cost of expression, and hence the strength of selection for specific activity, increases less rapidly with expression level than it would without selection for codon choice. When codon choice affects translational accuracy, the effect on the nonsynonymous rate is negligible. Because the translational error rate is always much smaller than one, the error rate has little effect on the consequences of a nonsynonymous change. As a result, the predicted nonsynonymous rates shown in figure 7*b* are in excellent agreement with the simulation results despite the fact that they neglect translational errors.

Drummond and Wilke (2008) considered the pairwise relationships among five variables: expression level; d*N* (proportional to the nonsynonymous rate); d*S* (proportional to the synonymous rate); the transition to transversion ratio, ts/tv; and the fraction of optimal codons, F_{op}. The directions of empirical correlations among these variables were reproduced by their simulations of a model of the MIM hypothesis. The simple binary model used here lacks a distinction between transitions and transversions. However, according to Drummond and Wilke (2008), ts/tv correlates with the other variables only because it is a proxy for the ratio of synonymous to nonsynonymous changes. Thus, if we substitute d*S*/d*N* for transition:transversion ratio, we can consider all ten pairwise relationships. As figure 8 shows, the model presented here produces relationships with exactly the same signs as those observed by Drummond and Wilke (2008), whether selection on codon choice is driven by translational efficiency or translational accuracy. Thus, selection for protein function can reproduce all ten of the pairwise relationships produced by simulations of a model of the MIM hypothesis and observed empirically.

The directions of these relationships can be understood in terms of the effect of expression level on each of the other four variables. That d*N* is a decreasing function of expression level is the main prediction of the model. If the strength of synonymous selection increases with expression level, as it does under the models considered here, then d*S* will decrease with expression level and F_{op} will increase. If d*N* is more sensitive to expression level than is d*S*, d*S*/d*N* will decrease with expression level, as it does in figure 8. The directions of all the relationships shown in figure 8 then follow. These relationships can likely be produced by a wide variety of models that predict a negative correlation between expression level and protein evolutionary rate.

Conclusions

The negative correlation between expression level and protein evolutionary rate can be explained by a model based on selection for protein function. This force has generally been considered to be the main constraint on protein evolution and must constrain protein sequences to some extent. The model can, like the MIM hypothesis, reproduce several empirically observed relationships. Selection for function should not be rejected as the main constraint on protein evolution or the dominant determinant of protein evolutionary rate.

Acknowledgments

I thank David Lipman, Scott Roy, and Yuri Wolf for comments on the manuscript. This research was supported by the Intramural Research Program of the National Institute of Health, National Library of Medicine.

Literature Cited

Akashi H. 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics 136: 927–935.

Cherry JL. 1998. Should we expect substitution rate to depend on population size? Genetics. 150:911–919.

Dekel E, Alon U. 2005. Optimality and evolutionary tuning of the expression level of a protein. Nature. 436:588–592.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 102:14338–14343.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol. 23:327–337.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell. 134:341–352.

Gottschalk G. 1986. Bacterial metabolism. 2nd ed. New York: Springer-Verlag. 359p.

Gout J, Kahn D, Duret L. 2010. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genetics. 6:e1000944.

Gwartney JD, Stroup RL, Sobel RS, MacPherson D. 2008. Economics: private and public choice. 12th ed. Cincinnati (OH): South Western College Pub. 802p.

Hartl DL, Dykhuizen DE, Dean AM. 1985. Limits of adaptation: the evolution of selective neutrality. Genetics. 111:655–674.

Hurst LD, Smith NG. 1999. Do essential genes evolve slowly? Curr Biol. 9:747–750.

Ikemura T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol. 151:389–409.

Kacser H, Burns JA. 1973. The control of flux. Symp Soc Exp Biol. 27:65–104.

Kimura M. 1957. Some problems of stochastic processes in genetics. Ann Math Statist. 28:882–901.

Kimura M. 1986. DNA and the neutral theory. Philos Trans R Soc Lond B Biol Sci. 312:343–354.

Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. 13:2229–2235.

Makalowski W, Boguski MS. 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc Natl Acad Sci U S A. 95:9407–9412.

Neidhardt FC, Ingraham JL, Schaechter M. 1990. Physiology of the bacterial cell: a molecular approach. Sunderland (MA): Sinauer Associates. 506p.

Oliphant TE. 2007. Python for scientific computing. Comput Sci Eng. 9:10–20.

Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics. 158:927–931.

Pál C, Papp B, Hurst LD. 2003. Genomic function: rate of evolution and gene dispensability. Nature. 421:496–497.

Rocha EP, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol. 21:108–116.

Sharp PM, Li WH. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol. 24:28–38.

Stein LD, et al. 2003. The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. PLoS Biol. 1:E45.

Stoebel DM, Dean AM, Dykhuizen DE. 2008. The cost of expression of *Escherichia coli* lac operon proteins is in the process, not in the products. Genetics. 178:1653–1660.

Waterston RH, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520–562.

Wolf YI, Gopich IV, Lipman DJ, Koonin EV. 2010. Relative contributions of intrinsic structural-functional constraints and translation rate to the evolution of protein-coding genes. Genome Biol Evol. 2:190–199.

Associate editor: Laurence Hurst