Relative Contributions of Intrinsic Structural–Functional Constraints and Translation Rate to the Evolution of Protein-Coding Genes

Yuri I. Wolf[1], Irina V. Gopich[2], David J. Lipman[1], and Eugene V. Koonin*[,1]

[1]National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland

[2]National Institute of Diabetes, Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland

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

Accepted: 8 March 2010

Abstract

A long-standing assumption in evolutionary biology is that the evolution rate of protein-coding genes depends, largely, on specific constraints that affect the function of the given protein. However, recent research in evolutionary systems biology revealed unexpected, significant correlations between evolution rate and characteristics of genes or proteins that are not directly related to specific protein functions, such as expression level and protein–protein interactions. The strongest connections were consistently detected between protein sequence evolution rate and the expression level of the respective gene. A recent genome-wide proteomic study revealed an extremely strong correlation between the abundances of orthologous proteins in distantly related animals, the nematode *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster.* We used the extensive protein abundance data from this study along with short-term evolutionary rates (ERs) of orthologous genes in nematodes and flies to estimate the relative contributions of structural–functional constraints and the translation rate to the evolution rate of protein-coding genes. Together the intrinsic constraints and translation rate account for approximately 50% of the variance of the ERs. The contribution of constraints is estimated to be 3- to 5-fold greater than the contribution of translation rate.

Key words: protein evolution, structural–functional constraints, misfolding, protein abundance.

Introduction

The rates of evolution of protein-coding genes span a range of three to four orders of magnitude but each gene has a characteristic rate that remains relatively constant over long evolutionary intervals (Zuckerkandl and Pauling 1965). Genome-wide measurements of evolutionary rates (ERs) revealed a remarkable constancy of the shape of the distributions of the rates across sets of orthologous genes in diverse life forms, from bacteria to mammals (Grishin et al. 2000; Wolf et al. 2009). The universality of the ER distribution implies simple and equally universal underlying determinants. The nature of these factors, arguably, is one of the central problems of evolutionary biology. It is traditionally assumed that ER is a multiplicative function of, first, the intrinsic structural–functional constraints that affect the given protein and, second, the biological importance of the protein in the organism (Wilson et al. 1977). Until recently, this hypothesis and the relative contributions of the two terms remained effectively inaccessible to empirical study.

Functional genomics and systems biology revealed a complex structure of correlations between evolutionary and phenomic variables (Herbeck and Wall 2005; Koonin and Wolf 2006; Pal et al. 2006; Vitkup et al. 2006; Wolf 2006) which comprise two distinct classes so that within-class correlations are positive whereas between-class correlations are negative (Wolf et al. 2006). For instance, the ER and propensity for gene loss are positively correlated; by contrast, each of these variables is negatively correlated with the gene expression level. Surprisingly, little if any correlation was detected between the essentiality of genes for the reproduction of organisms and the ER: at best, nonessential genes evolve slightly faster than essential genes (Hurst and Smith 1999; Hirsh and Fraser 2001; Jordan et al. 2002; Krylov et al. 2003; Wall et al. 2005; Wolf 2006). Among all the detected connections, the most consistent

and strongest one is the negative correlation between the expression level of a gene and its sequence evolution rate: highly expressed genes evolve significantly slower than lowly expressed ones (Pal et al. 2001; Krylov et al. 2003; Drummond et al. 2005; Lemos et al. 2005).

The link between expression level and sequence evolution is invariably detected across a broad range of model organisms, so it was proposed that expression level or, more precisely, the rate of translational events is the dominant determinant of the sequence evolution rate (Drummond et al. 2005, 2006; Drummond and Wilke 2008). This idea is embodied in the mistranslation-induced misfolding (MIM) hypothesis according to which the underlying cause of the covariation between the sequence evolution rate and expression level is the selection for robustness to protein misfolding, that is, increasingly important for highly expressed genes owing to the toxic effects of misfolded proteins (Drummond et al. 2006; Wilke and Drummond 2006; Drummond and Wilke 2008, 2009). Detailed computer simulations of protein evolution seem to indicate that the toxic effect of protein misfolding, indeed, could suffice to explain the observed covariation of expression level and sequence evolution rate (Drummond and Wilke 2008). An empirical test of the MIM hypothesis indicated that the ERs of domains in multidomain proteins (in which the domains are translated at the same rate) are substantially homogenized compared with the ERs of the same domains in separate proteins (Wolf et al. 2008). This observation directly supports the hypothesis that the translation rate is one of the determinants of protein evolution and suggests that the contribution of this factor might be comparable with that of structural–functional constraints.

A recent comparative proteomic study of two distantly related model animals, the nematode *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster*, revealed an unexpectedly strong positive correlation (correlation coefficient of ~0.8) between the abundances of orthologous proteins in the two organisms (Schrimpf et al. 2009). This finding seems to be compatible with the generalized MIM hypothesis because protein abundance comes across as an evolutionarily highly conserved and, by implication, critically important feature. Although protein abundance is obviously a function of both translation rate and protein degradation rate, experimental studies suggest that the contribution of translation rate is much greater than that of the degradation rate (Belle et al. 2006).

Here, we show that the correlation between protein abundances is much higher than the correlation between the ERs of orthologous genes in the nematode and the fly. We then develop a mathematical model that allows a quantitative estimation of the relative contributions of structural–functional constraints and translation rate to the rate of evolution of protein-coding genes and shows that structural–functional constraints are the primary factor shaping protein evolution.

Materials and Methods

Genome sequences of *D. melanogaster* were obtained from the FlyBase database. Genome sequences of *Drosophila pseudoobscura*, *C. elegans*, and *Caenorhabditis briggsae* were obtained from the National Center for Biotechnology Information's RefSeq database (Wheeler et al. 2003). Reciprocal BlastP (Altschul et al. 1997) searches (*e* value threshold 1×10^{-6}, effective database size 2×10^{7}, no low-complexity filtering or composition-based statistics) were performed for *C. elegans*–*C. briggsae* and *D. melanogaster*–*D. pseudoobscura* genome pairs. Putative orthologs were identified as bidirectional best hits (Tatusov et al. 1997). Protein sequences of orthologs were aligned using MUSCLE (Edgar 2004). Lineage-specific ERs were estimated as amino acid distances between aligned sequences of orthologs and were calculated using PROTDIST (Felsenstein 1996) with the Jones-Taylor-Thornton evolutionary model (Jones et al. 1992), and gamma-distributed site rates with the shape parameter equal to 1.0. If the amino acid sequences of orthologs were identical, a distance of 0.5/length was assigned.

Protein and mRNA abundance data for the nematode *C. elegans* and the fruit fly *D. melanogaster* (Schrimpf et al. 2009) and the worm-fly orthology relationship data were kindly provided by Manuel Weiss and Sabine Schrimpf (University of Zurich, Zurich, Switzerland). When the orthology relationship involved multiple genes from one of both organisms, the most similar pair was included (a simplified Index Ortholog procedure Krylov et al. 2003; Wolf et al. 2006).

Assignments of worm and fly genes to EggNOGs (Jensen et al. 2008) were used to ascribe a functional class (Tatusov et al. 2003) to a worm-fly pair of orthologs.

ER and mRNA abundance data for human and mouse proteins were from Wolf et al. (2009).

Logarithms of protein (mRNA) abundances and evolution rates for 2,297 quartets of orthologs were standardized to the average of 0 and standard deviation of 1.

Results and Discussion

Correlations between Evolutionary and Phenomic Variables

Considering the unexpected high correlation between the abundances of orthologous proteins in the nematode and the fly (Schrimpf et al. 2009), we reexamined the data and compared this correlation with the correlation between the rates of sequence evolution among orthologous genes in the respective lineages. To this end, we calculated lineage-specific, short-term ERs by comparing the sequences of orthologous genes for the nematodes *C. elegans* and *C. briggsae* and the flies *D. melanogaster* and *D. pseudoobscura*. The two pairs of species are separated by nearly the same evolutionary distance and show nearly

Table 1
Measured Correlations between Protein Abundances and Lineage-Specific Evolutionary Rates

Variable	Nematode	Fly
r_A	+0.80	
r_R	+0.52	
r_{RAxx}, r_{RAYY}	−0.41	−0.34
r_{RAxY}, r_{RAYx}	−0.37	−0.32
r_Δ	−0.09*	

*$P = 1.7 \times 10^{-5}$

identical distributions of ERs between orthologs (supplementary fig. S1, Supplementary Material online). Altogether we identified 2,297 quartets of orthologs for which reliable abundance data (Schrimpf et al. 2009) were available as well. The correlations between protein abundances and lineage-specific ERs within this set of orthologs are shown in figure 1*A* and 1*B*, and table 1. The correlation between protein abundances was nearly identical to the value reported by Schrimpf et al. (2009), whereas the correlation between the ERs was substantially lower (~0.52 for the rates vs. ~0.80 for the abundances; compare figures 1*A* and 1*B*). For each lineage, a moderate but highly significant negative correlation was observed between the ER and protein abundance (fig. 1*C* and 1*D*), in agreement with the universal negative correlation between ER and expression level (Pal et al. 2001; Krylov et al. 2003; Wolf et al. 2006; Drummond and Wilke 2008, 2009).

Thus, protein abundance seems to be controlled by purifying selection much more tightly than the ER, regardless of the factors that determine the latter (provided that the measurement noise is of comparable magnitude for both variables). The availability of two parallel arrays of ER and protein abundance data for orthologs from distantly related animals prompted us to attempt to disentangle the contributions of structural–functional constraints and translation rate to the ER. Orthologs from different animals are highly similar structurally and functionally, so to a good approximation the structural–functional constraints can be assumed to be the same. Under this assumption, although the correlations between the ER and the abundances of worm and fly proteins, considered separately, are affected by both structural–functional and translation rate–determined effects, the correlations between the differences in the ER and the differences between the protein abundances in two organisms (hereinafter r_Δ) should be determined solely by the translation rates and random noise.

Hence an important reality check: if the observed difference between abundances of orthologous proteins is biologically relevant rather than caused by random noise, r_Δ should have the correct sign (same as in the rate-abundance correlation) and be statistically significant. Our calculation yielded $r_\Delta \approx -0.09$ (table 1, fig. 1*E*), a relatively low but statistically significant value ($P = 1.7 \times 10^{-5}$). Moreover, estimates of r_Δ were consistent with respect to the sign, magnitude, and statistical significance of the correlation when different, independent data sets were analyzed and were supported by a bootstrap test (see below). Thus, we proceed with a formal model of the effects of constraints and translation rate on the ERs and solve this model for its parameters.

Modeling Evolution of Protein-Coding Genes to Infer the Relative Contributions of Structural–Functional Constraints and Translation Rate

Assumptions. We developed a mathematical model to use the data on the correlations between protein abundances and ERs in two lineages to infer the relative contributions of constraints and translation rates to the evolution of protein-coding genes. The model rests on the following assumptions.

1. ER can be broken down into a product of the following components:
   - o translation rate-dependent factors;
   - o factors that are independent of translation rate but are common for orthologs in the compared organisms, and
   - o other factors that are independent of translation rate and independent between orthologs;

We refer to the factors that are common for orthologs but independent of the translation rate as "structural–functional constraints." This interpretation appears plausible because orthologous protein, at least, those that show high sequence conservation, typically possess the same overall structure, retain the same function and operate in similar cellular contexts, even in organisms separated by hundreds of millions of years of evolution, such as representatives of different animal phyla. Both the intrinsic constraint term and the translation rate-dependent term can be approximated by power functions (linear functions of log variables in the log scale) of a protein-specific "structural–functional factor" and translation rate, respectively.

2. There is a substantial component in the contribution of structural–functional constraints, that is, lineage independent and gene independent.
3. There is a substantial component in the contribution of translation rate, that is, gene independent but lineage specific.

Thus, the above two assumptions refer to the genome-wide factors that determine the relationships between structural–functional constraints and effects of translation rate. In addition, each gene has a specific translation rate and unique structural–functional constraints that affect

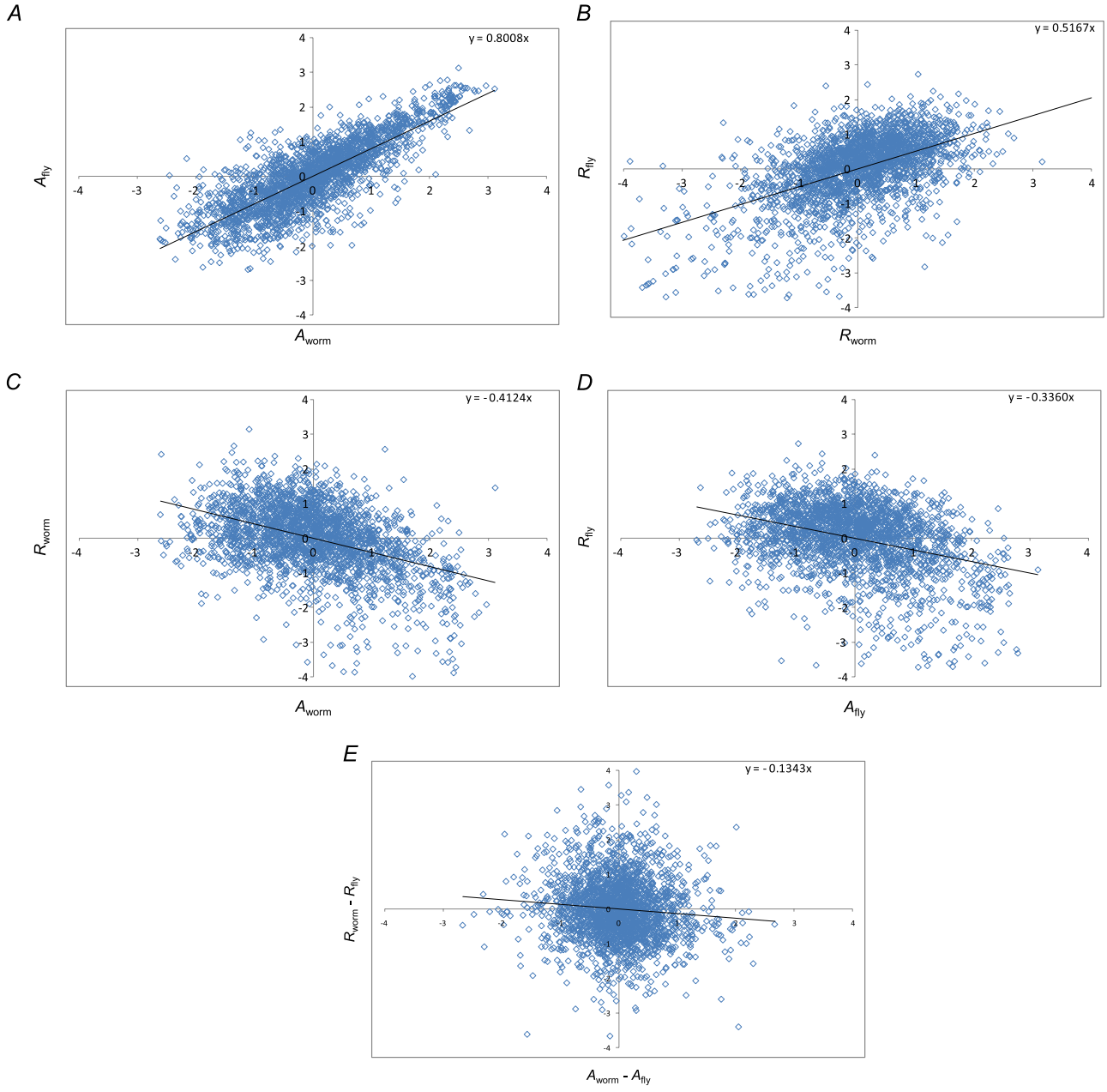

Fig. 1.—Correlations between abundances and evolutionary rates of orthologous proteins in nematodes and flies. *A*) Protein abundances in *C. elegans* and *D. melanogaster*. *B*) Evolutionary rates in the nematode and fly lineages. *C*) Protein abundance versus evolutionary rate in the nematode. *D*) Protein abundance versus evolutionary rate in the fly. *E*) Difference in abundances versus difference in evolution rates.

the relative contribution of these factors to the evolution rate of this gene.

4. Translation rate is approximated by the abundance of the corresponding gene product (protein or mRNA); the difference between protein (mRNA) abundances is negligible between closely related species but substantial between distantly related lineages (this assumption is compatible with the results of genome-wide studies on evolution of gene expression in diverse model organisms Jordan et al. 2005; Khaitovich et al. 2006); the error of the abundance estimate is independent of other variables but could be correlated between orthologs in different lineages.
5. Effects of other translation-independent factors that differ between orthologs in different organisms, random noise, and errors of rate measurement can be combined into a single variable which is independent of other variables.
6. Means and variances of the distributions of all variables are finite.

The Model. With the above assumptions, for gene *i* in species *X* and *Y*, respectively, the ER (on the log scale) is:

$$R_{X,i} = \beta S_i + \alpha_X T_{X,i} + E_{X,i},$$
$$R_{Y,i} = \beta S_i + \alpha_Y T_{Y,i} + E_{Y,i}, \quad (1)$$

where $R_{x,i}$ is the ER of gene *i* in the lineage *X*, S_i is the gene-specific constraint factor assumed to be identical for orthologous genes in the two lineages, $T_{x,i}$ is the translation rate of gene *i*, $E_{x,i}$ is the gene-specific combination of random and unknown factors, and α_x and β are coefficients that reflect the gene-independent (genome-wide) components of the relative contributions of the constraints and translation rate, respectively (same for gene *i* in species *Y*).

In practice, the translation rates cannot be measured directly but are correlated with the observable abundances of gene products:

$$A_{X,i} = \gamma T_{X,i} + \epsilon_{X,i},$$
$$A_{Y,i} = \gamma T_{Y,i} + \epsilon_{Y,i}, \quad (2)$$

where $A_{x,i}$ is the observed abundance of the *i*-th gene product in species *X*, γ is the "accuracy coefficient" that reflects the correlation between abundance and the actual (hidden) translation rate, and $\epsilon_{x,i}$ is the component of the observed abundance that encompasses gene-specific measurement errors and other random factors (same for the *i*-th gene in the species *Y*).

If ER, translation rates, abundances, and constraint factors each are standardized on the log scale to the mean of 0 and variance of 1, then:

$$\langle R_{X,i}^2\rangle = \langle R_{Y,i}^2\rangle = \langle T_{X,i}^2\rangle = \langle T_{Y,i}^2\rangle = \langle A_{X,i}^2\rangle = \langle A_{Y,i}^2\rangle = \langle S_i^2\rangle = 1$$

(where $\langle a_i\rangle$denotes the expectation of a_i across all *i*). The fraction of the total variance of *R* unexplained by *S* and *T* ($\langle E_{X,i}^2\rangle$ and$\langle E_{Y,i}^2\rangle$) is unknown, whereas $\langle \epsilon_{X,i}^2\rangle = \langle \epsilon_{Y,i}^2\rangle = 1 - \gamma^2$ (from eq. (2)). As random factors are uncorrelated with each other or with other variables, expectations of all cross products involving *E* or ϵ are equal to zero with the exception of $\langle \epsilon_{X,i}\epsilon_{Y,i}\rangle$ (see below).

Solution of the Model. From the equations (1) and (2):

$$\begin{aligned}
r_R &= \langle R_{X,i}R_{Y,i}\rangle = \beta^2 + \alpha_X\alpha_Y\langle T_{X,i}T_{Y,i}\rangle \\
&\quad + \alpha_X\beta\langle S_iT_{X,i}\rangle + \alpha_Y\beta\langle S_iT_{Y,i}\rangle, \\
r_A &= \langle A_{X,i}A_{Y,i}\rangle = \gamma^2\langle T_{X,i}T_{Y,i}\rangle + \langle \epsilon_{X,i}\epsilon_{Y,i}\rangle, \\
r_{RAXX} &= \langle R_{X,i}A_{X,i}\rangle = \gamma(\alpha_X + \beta\langle S_iT_{X,i}\rangle), \\
r_{RAYY} &= \langle R_{Y,i}A_{Y,i}\rangle = \gamma(\alpha_Y + \beta\langle S_iT_{Y,i}\rangle), \\
r_{RAXY} &= \langle R_{X,i}A_{Y,i}\rangle = \gamma(\alpha_X\langle T_{X,i}T_{Y,i}\rangle + \beta\langle S_iT_{Y,i}\rangle), \\
r_{RAYX} &= \langle R_{Y,i}A_{X,i}\rangle = \gamma(\alpha_Y\langle T_{X,i}T_{Y,i}\rangle + \beta\langle S_iT_{X,i}\rangle)
\end{aligned} \quad (3)$$

(the names for the correlations that can be measured from the data are assigned for convenience). Additionally, we express the correlation between the deviations of the experimentally measured abundances from the true translation rates using the correlation coefficient r_e $\langle \epsilon_{X,i}\epsilon_{Y,i}\rangle = r_e(1-\gamma^2)$. Then the system (3) can be solved with respect to α_x, α_Y and β using γ and r_e as free parameters:

$$\begin{aligned}
\alpha_X &= \frac{r_{RAXX} - r_{RAYX} + (r_{RAYY} - r_{RAXY})\langle T_{X,i}T_{Y,i}\rangle}{\gamma(1-\langle T_{X,i}T_{Y,i}\rangle^2)}, \\
\alpha_Y &= \frac{r_{RAYY} - r_{RAXY} + (r_{RAXX} - r_{RAYX})\langle T_{X,i}T_{Y,i}\rangle}{\gamma(1-\langle T_{X,i}T_{Y,i}\rangle^2)}, \\
\beta^2 &= r_R - \alpha_X\alpha_Y\langle T_{X,i}T_{Y,i}\rangle \\
&\quad - \alpha_X\left(\frac{r_{RAXX}}{\gamma} - \alpha_X\right) - \alpha_Y\left(\frac{r_{RAYY}}{\gamma} - \alpha_Y\right),
\end{aligned} \quad (4)$$

where $\langle T_{X,i}T_{Y,i}\rangle = \frac{r_A - r_e(1-\gamma^2)}{\gamma^2}$. Additionally,

$$\begin{aligned}
\langle E_{X,i}^2\rangle &= 1 - \beta^2 - \alpha_X^2 - 2\alpha_X\left(\frac{r_{RAXX}}{\gamma} - \alpha_X\right), \\
\langle E_{Y,i}^2\rangle &= 1 - \beta^2 - \alpha_Y^2 - 2\alpha_Y\left(\frac{r_{RAYY}}{\gamma} - \alpha_Y\right), \\
\langle S_iT_{X,i}\rangle &= \left(\frac{r_{RAXX}}{\gamma} - \alpha_X\right)/\beta, \\
\langle S_iT_{Y,i}\rangle &= \left(\frac{r_{RAYY}}{\gamma} - \alpha_Y\right)/\beta, \\
r_\Delta &= \frac{r_{RAXX} + r_{RAYY} - r_{RAXY} - r_{RAYX}}{2\sqrt{(1-r_R)(1-r_A)}}.
\end{aligned} \quad (5)$$

Equation (4) gives the absolute value for β without any indication of its sign. Indeed, given that here *S* is a hidden, not directly observable variable, it can be construed as either a measure of constraint (negatively correlated with *R*) or as a measure of robustness to mutational and translational errors (positively correlated with *R*). Hereinafter, we interpret *S* as a constraint and, accordingly, assume β to be negative.

Exploring the Parameter Space. Equations (4–5) allow one to estimate the relative contributions of intrinsic constraints and translation rate to the ER of protein-coding genes from the correlations between the variables (eq. (3), table 1) if the accuracy coefficients γ connecting the observed gene product abundance with the hidden translation rate and r_e, the correlation between the abundance measurement errors, are known. The available data do not allow a direct estimate for γ and r_e but several observations can be made regarding these parameters.

Both γ and r_e are correlation coefficients, the former between the translation rate and measured abundance (eq. (2)) and the latter between measurement errors for orthologs in different organisms. Thus, both must be less than or equal to 1. Moreover, the values of γ and r_e have to conform to several boundary conditions arising from the nature of the model variables and parameters (supplementary table S1, Supplementary Material online); for instance, we expect a nonnegative correlation between the measured and real values.

Estimation of the Relative Contributions of Structural–Functional Constraints and Translation Rate to Protein Evolution. The case of $\gamma \to 1$ (the value of r_e becomes irrelevant here) implies a perfect correspondence

Table 2
Estimated Model Parameters

Variable	$\gamma = 1$, $r_e = 0$	Median	Source
$\langle T_{X,i}T_{Y,i}\rangle$	+0.80	+0.82	equation (4)
α_x, α_Y	−0.17, −0.10	−0.22, −0.13	equation (4)
β	−0.68	−0.64	equation (4)
β/α	4.0, 6.9	2.9, 4.9	
$\langle S_iT_{X,i}\rangle$, $\langle S_iT_{Y,i}\rangle$	+0.36, +0.35	+0.37, +0.37	equation (5)
$\langle E^2_{X,i}\rangle$, $\langle E^2_{Y,i}\rangle$	0.43, 0.48	0.43, 0.51	equation (5)

between translation rate and the measured protein abundance. Under this assumption, we estimate the β/α ratio to be in the range of 4–7 and the correlations between the constraint and translation rate factors ($\langle S_iT_{X,i}\rangle$ and $\langle S_iT_{Y,i}\rangle$) in the range of 0.35-0.36 depending on organism (table 2).

Perhaps, not surprisingly, all boundary conditions (supplementary table S1, Supplementary Material online) combined exclude more than 3/4 of the possible values of γ and r_e (fig. 2*A*). Numerical exploration of the (γ, r_e) parameter space (fig. 2 and supplementary fig. S3B–K, Supplementary Material online) reveals a singularity area where the absolute values of α_x, α_Y, and β increase above 1, the β/α ratio drops to ~0.8, the correlation between organism-specific translation rates $\langle T_{X,i}T_{Y,i}\rangle$ approaches 1, the correlations between the constraint factors and translation rates ($\langle S_iT_{X,i}\rangle$ and $\langle S_iT_{Y,i}\rangle$) approach −1, and the residual fraction of the variance of R ($\langle E^2_{X,i}\rangle$, $\langle E^2_{Y,i}\rangle$) declines toward 0. This area corresponds to unrealistic relationships between the ER, structure-functional constraints, and translation, with virtually no real differences between orthologs ($\langle T_{X,i}T_{Y,i}\rangle \rightarrow 1$) but with amplification of whatever tiny fraction of variance in $T_{x,i}$–$T_{Y,i}$ remains by the very high absolute values of α_x and α_Y (extremely strong amplification by translation). Because the variance of R is assumed to be equal to 1, this dictates comparably high absolute values of β and a very strong negative correlation between S and T.

In the absence of any reliable a priori information about γ and r_e, we assume a uniform distribution of these parameters within the range of γ and r_e that is compatible with the boundary conditions (fig. 2*A*). Then, we can estimate median values for all parameters and variables, that is, values such that, for half of the area within the domain, the surface of the corresponding function lies below and for the other half above this value (table 2). Given the relative flatness of the surfaces representing the parameter values (fig. 2*B,C,D*), medians seem to be a good representation of the values "typical" for the system.

The relative contribution of the structural–functional constraints to the ER is predicted to be greater than the contribution of the translation rate ($\beta/\alpha > 1$) over most of the parameter space with the exception of the neighborhood of the singularity (fig. 2). Using the median as a realistic middle ground, we find that the effect of constraints is approximately 3- to 5-fold greater than the effect of the translation rate. The fraction of the variance of the ER, that is, explained by the combination of the constraint and translation rate factors ($1 - \langle E^2_{X,i}\rangle$, $1 - \langle E^2_{Y,i}\rangle$) remains remarkably stable at ~50%.

Protein abundance data can explain 10–17% (r^2_{RA}) of the ER variance within an organism. In part, this contribution seems to arise from a joint effect of structural–functional constraints and translation rate as there is a moderate but substantial positive correlation between S and T (median values +0.37 for both organisms). These findings suggest that structural–functional constraints could partly determine the allowable abundance of proteins. Again, using the median values, we estimate that translation rate alone would explain only 2–5% (α^2) of the original variance of ER, whereas structural–functional constraints, if amenable to direct measurement, alone would explain ~41% (β^2) of the original variance; the remaining 8–15% of the variance is explained by the joint contribution of the constraint and translation rate factors (fig. 3).

To assess the robustness of the above estimates to sampling bias, we used two approaches. First, we produced 1,000 bootstrap replications of the pairs of orthologous genes, computed the correlations for the bootstrapped samples, and estimated the model parameters for each sample (supplementary table S2A, Supplementary Material online; the estimates are for $\gamma \rightarrow 1$). The 95% confidence intervals for the β/α ratio in the resampled data were 2.8–6.2 and 3.9–20.7 for β/α_x and β/α_Y, respectively (compare with the values in table 2). Second, we analyzed four broad functional classes of genes (information storage and processing, cellular processes and signaling, metabolism, and poorly characterized Tatusov et al. 2003) separately (supplementary table S2B, Supplementary Material online; the estimates are for $\gamma \rightarrow 1$). Due to an approximately 4-fold reduction of the sample size, neither of these categories gives a statistically significant r_Δ value. Nevertheless, the estimates for the β/α ratio stay within the same range (1.7–18.6, supplementary table S2B, Supplementary Material online) across all the classes. Thus, the results are robust to sampling error and do not depend on the presence of a small number of biased sets of orthologs.

The same approach to modeling evolution of protein-coding genes can be implemented also by using the mRNA abundance data as a proxy for the translation rate. As noticed by Schrimpf et al. (2009), mRNA abundance data are relatively poorly correlated between nematodes and flies and with the ERs (supplementary table S3, Supplementary Material online) compared with the protein abundance data (table 1). Nevertheless, the r_Δ value computed for the mRNA abundance data remained significant ($r_\Delta = -0.05$, $P = 1.1 \times 10^{-2}$), so we used it to perform the same calculations (supplementary table S4, Supplementary Material online). The estimate range for the median β/α ratio (5–25) was generally consistent with the values obtained for the protein abundance data. The

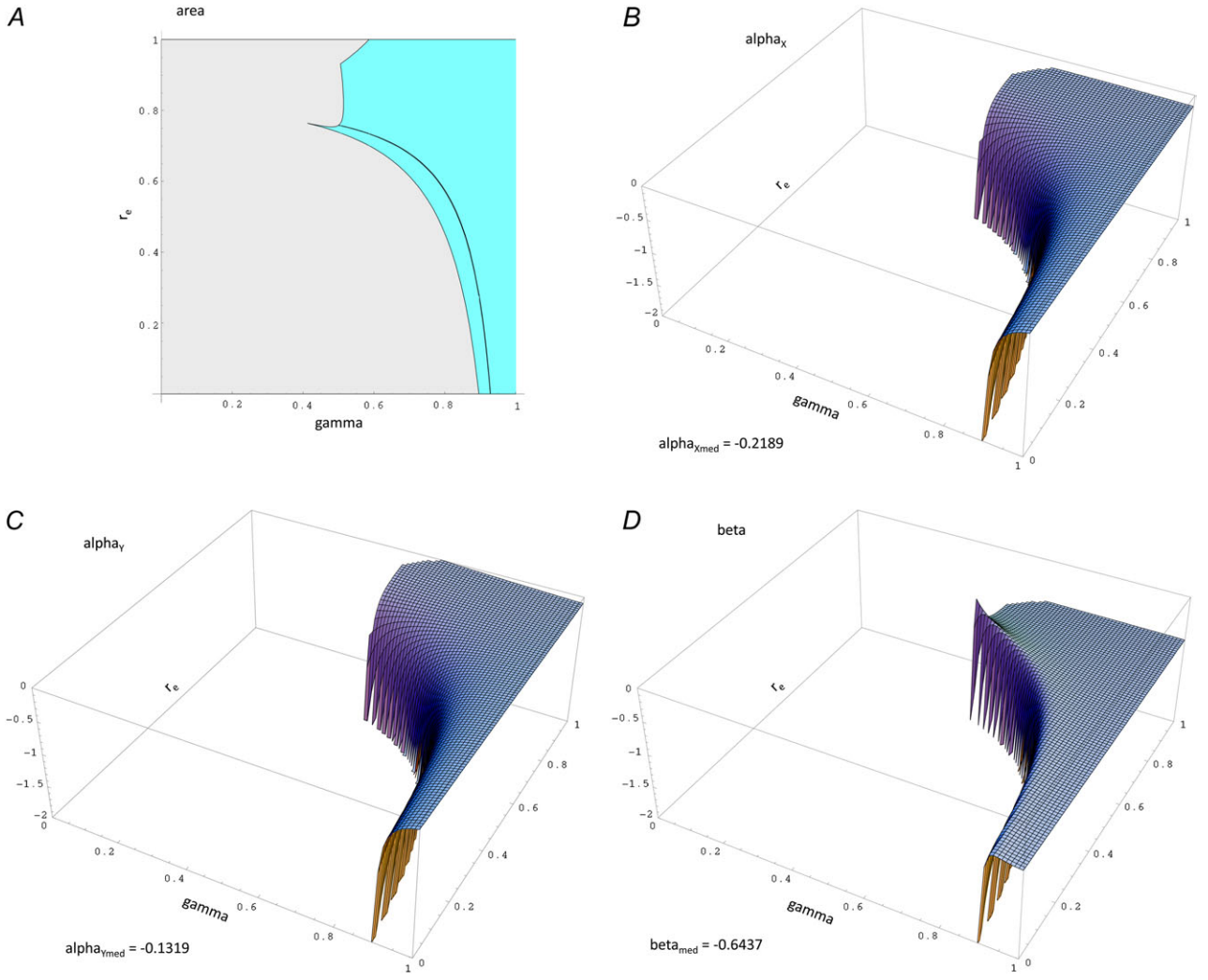

Fig. 2.—Relationships between the model parameters γ and r_e and the key variables. *A*) Area of the parameter space satisfying the boundary conditions from Table 2. *B*) Values of α_X. *C*) Values of α_Y. *D*) Values of β.

wider range of values, probably, is caused by the relatively low correlations between mRNA abundances. Similar results, albeit with an even greater scatter (the β/α ratio in the range of 7–50), were obtained for 8,511 human–mouse orthologs using expressed sequence tag counts as a proxy for expression level (supplementary table S5, Supplementary Material online).

The estimations of the relative contributions of structural–functional constraints and translation rate (abundance) to the evolution of protein-coding genes critically depend on the use of r_Δ, which is a small value. This could be an issue of concern but the fact that r_Δ is statistically significant for the independently measured protein and mRNA abundances and that estimations using both data sets yield compatible values of the β/α ratio suggests that the resulting estimates are valid and reasonably robust.

For the final and, arguably, crucial test of the above conclusions, we employed the data on protein abundance of proteins in the plant *Arabidopsis thaliana* (kindly provided by Christian von Mering) paired with the fruit fly data to repeat the estimation of the relative contributions of the constraint and translation rate factors to the ER (supplementary table S6, Supplementary Material online). The resulting ratios of the medians of β and α are in the range of 1.6–7.1, in a good agreement with the results obtained with the fly and nematode data. The congruence of the results obtained with organisms as evolutionarily distant as animals and plants suggests that the relative contributions of the constraint and translation rate factors to protein evolution could be universal across the entire diversity of cellular life forms.

A General Model of Misfolding-Driven Protein Evolution

The MIM hypothesis postulates the central role of MIM cost in determining the selection pressure experienced by a protein-coding gene (Drummond et al. 2006; Drummond and

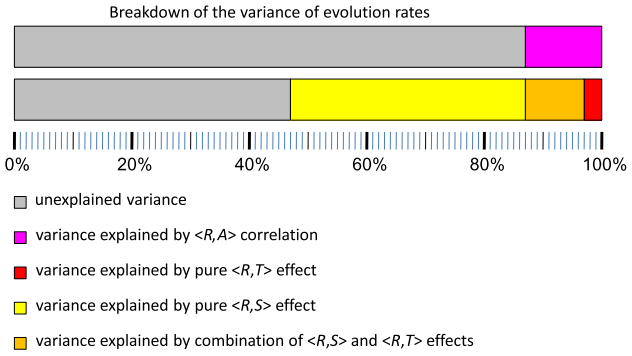

Fig. 3.—Relative contributions of structural–functional constraints and protein abundance (translation rate) to the evolution of protein-coding genes. Top: accounting for protein abundance; bottom: accounting for translation rate and structural–functional constraints.

Wilke 2008). Our present results suggest an even more general model to explain the dependence of protein evolution on both the intrinsic structural–functional constraints and translation rate. A protein fold or (super)family can be represented as a peak in a protein folding landscape where the plane corresponds to the sequence space and the altitude is the probability that a given sequence will fold correctly, hereinafter "robustness" (fig. 4). Under the MIM hypothesis, the altitude of a point in the sequence space depends not only on the robustness of the exact replica of the corresponding protein but also on the robustness of each of the mistranslation products of the mRNA coding for this protein, weighted by the probability of emergence of a particular mistranslated variant. The total cost of misfolding for a given protein is determined by the amount of misfolded forms; this amount is proportional to the translation rate and inversely proportional to misfolding robustness. More precisely, the fitness difference between two alleles depends (possibly, in a nonlinear fashion) on the difference between misfolding costs incurred by the expression of these alleles.

Selection to reduce the misfolding cost favors mutations that increase robustness, whereas random drift tends to scatter protein sequences away from the summit and down the slope of a robustness peak. These two trends reach an equilibrium at some cost level; the corresponding equilibrium level of robustness depends on the translation rate: highly expressed proteins must be highly robust, otherwise the misfolding cost would be unacceptably high (fig. 4). This model yields a possible explanation of the apparently paradoxical observation that, although highly expressed proteins are selected for higher robustness, they also are more constrained: the higher the equilibrium robustness level, the smaller the fraction of mutations that do not push protein robustness below this threshold (fig. 4). The model implies that both MIM and native sequence misfolding are important determinants of protein evolution, and the contribution of native sequence misfolding is the greatest for highly expressed proteins that have small robust sequence neighborhoods (fig. 4).

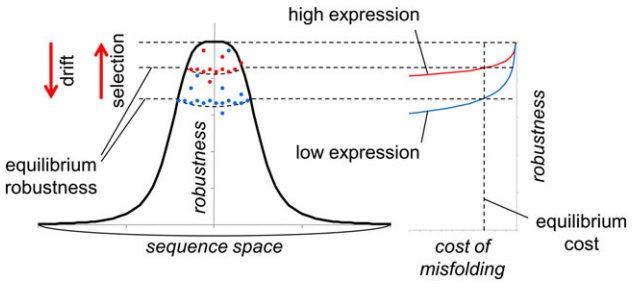

Fig. 4.—The general model of misfolding-driven protein evolution. The schematic shows the relationships between misfolding robustness, fitness, expression, selection, and drift in protein evolution.

The MIM hypothesis assumes, explicitly (Wilke and Drummond 2006) or implicitly (Drummond and Wilke 2008), that, although robustness peaks differ in height and shape between protein folds and (super)families, and in particular, robust folders have higher amino acid residue contact densities than less robust ones (Drummond and Wilke 2008; Zhou et al. 2008), these differences are less consequential than direct effects of translation rates. Our previous work showed that the effects of structural–functional constraints and translation rate are comparable (Wolf et al. 2008). The present findings that result from a completely different approach further extend and specify these conclusions, suggesting that the intrinsic difference in robustness between protein domains is the primary determinant of the ER, whereas translation rate alone explains but a small fraction of the variance (fig. 3). The positive correlation between the apparent pressure of structural–functional constraints and translation rate further implies that, although highly expressed proteins are likely to be more robust to misfolding than lowly expressed proteins, as a result of adaptation, the fitness landscape becomes increasingly rugged, with steeper peaks, as altitude (i.e., intrinsic misfolding robustness of the native sequence) increases. Thus, proteins that are highly robust to misfolding are conversely weakly robust to mutation as sequences in their immediate neighborhoods are substantially less robust to misfolding.

In principle, interpretation of the present results in terms of the robustness of proteins to misfolding is not strictly necessary. One could view the high contribution of the factor denoted *S* in our model as a measure of the "functional density" of a protein (Wilson et al. 1977). However, in contrast to the misfolding-rate hypothesis discussed above, the functional density perspective does not imply any physical mechanism to explain the universal dependence between evolution rate and the abundance of proteins. Furthermore, are misfolding-rate concept is compatible with the recent results on the connection between protein folding and evolution which indicate that the characteristic distribution of sequence evolution rate is a consequence of the fundamental physical principles of folding (Lobkovsky et al. 2010).

It should be stressed that the present results and conclusions are at no discrepancy with the widely supported observation that expression level of protein-coding genes is the best known predictor of the ER (Pal et al. 2001; Drummond et al. 2005, 2006; Wolf et al. 2006; Drummond and Wilke 2009). Instead, the results of this work provide a step toward dissection of this phenomenological connection into specific, mechanistic components, and suggest that the factor primarily responsible for the observed anticorrelation between expression level and ER is the intrinsic robustness of proteins to misfolding. The negative correlation between expression level (abundance) and ER comes across as the strongest because the even stronger relationship between intrinsic structural–functional constraints and ER (fig. 3) is not directly measurable (at least not without much extra effort). The present results do not invalidate the central point of the MIM hypothesis, that the cost of misfolding is a key determinant of protein evolution. However, our observations shift the emphasis from translation rate per se to intrinsic structural–functional constraints that in turn affect the translation rate and thus take the entire concept of misfolding-driven protein evolution closer to a specific, mechanistic model.

Concluding Remarks

The recently reported high-quality proteomic data for two distantly related animals (Schrimpf et al. 2009) followed by similar results for even more distantly related organisms, reveal not only a strong correlation between abundances of orthologous proteins in different organisms but also a relatively high correlation between protein abundances and evolution rates. We used these data to reexamine the determinants of the ERs of protein-coding genes. In a previous study, we showed that both intrinsic structural–functional constraints and the rate of expression made substantial and apparently independent contributions to the ER (Wolf et al. 2008). Here, we describe a mathematical model that takes advantage of the availability of comparative data on ER and protein abundances for several diverse lineages of eukaryotes to disentangle the contributions of the constraint factor and the translation rate factor and assess them quantitatively. We found that together, the two factors account for approximately 50% of the variance of the ER of proteins and that the contribution of structural–functional constraints is several-fold greater than the contribution of translation rate. Furthermore, the two factors are connected so that a protein's robustness to misfolding dependent on structural–functional constraints, to a large extent, determines the maximum allowable translation rate of the given protein.

The conclusions derived in this work directly apply only to subsets of proteins in each of the studied eukaryotic organisms that are, first, highly conserved in evolution so that orthologs between distant organisms can be identified with confidence and, second, are highly expressed so that they can be confidently identified by proteomic methods. These are "high status" (Wolf et al. 2006), largely house-keeping genes. Furthermore, comprehensive studies with different approaches and improved proteomic techniques should determine how general are the present conclusions on the relative roles of different factors in protein evolution.

The present model is based on several assumptions on the relationships between the key variables that affect evolution of protein-coding genes. Although these assumptions appear plausible, it would be important to investigate the possible effects of their violation. The limited amount of high-quality data on protein abundance presently does not allow us to investigate the full range of parameters. However, comprehensive analysis including validation of the present assumptions should become possible when such data become available for a wider range of organisms separated by a broader range of evolutionary distances.

Supplementary Material

Supplementary figures S1–S3 and supplementary tables S1–S6 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

We thank Sabine Schrimpf, Christian Von Mering, and Manuel Weiss (University of Zurich, Zurich, Switzerland) for providing the protein abundance data including an unpublished subset, and Josh Cherry, Alexander Lobkovsky, Scott Roy (National Center for Biotechnology Information), and Claus Wilke (University of Texas-Austin) for useful discussions. The authors' research is supported by the Department of Health and Human Services intramural funds (National Library of Medicine and National Institute of Diabetes, Digestive and Kidney Diseases, National Institutes of Health).

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Belle A, Tanay A, Bitincka L, Shamir R, O'Shea EK. 2006. Quantification of protein half-lives in the budding yeast proteome. Proc Natl Acad Sci USA. 103:13004–13009.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci USA. 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.

Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. Nat Rev Genet. 10:715–724.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
Felsenstein J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. 266:418–427.
Grishin NV, Wolf YI, Koonin EV. 2000. From complete genomes to measures of substitution rate variability within and between proteins. Genome Res. 10:991–1000.
Herbeck JT, Wall DP. 2005. Converging on a general model of protein evolution. Trends Biotechnol. 23:485–487.
Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. Nature 411:1046–1049.
Hurst LD, Smith NG. 1999. Do essential genes evolve slowly? Curr Biol. 9:747–750.
Jensen LJ, et al. 2008. eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res. 36:D250–D254.
Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275–282.
Jordan IK, Marino-Ramirez L, Koonin EV. 2005. Evolutionary significance of gene expression divergence. Gene. 345:119–126.
Jordan IK, Rogozin IB, Wolf YI, Koonin EV. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 12:962–968.
Khaitovich P, Enard W, Lachmann M, Paabo S. 2006. Evolution of primate gene expression. Nat Rev Genet. 7:693–702.
Koonin EV, Wolf YI. 2006. Evolutionary systems biology: links between gene evolution and function. Curr Opin Biotechnol. 17:481–487.
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.
Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol. 22:1345–1354.
Lobkovsky AE, Wolf YI, Koonin EV. 2010. Universal distribution of protein evolution rates as a consequence of protein folding physics. Proc Natl Acad Sci USA. doi:10.1073/pnas.0910445107.
Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics 158:927–931.
Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet. 7:337–348.
Schrimpf SP, et al. 2009. Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. PLoS Biol. 7:e48.
Tatusov RL, et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 4:41.
Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science. 278:631–637.
Vitkup D, Kharchenko P, Wagner A. 2006. Influence of metabolic network structure and function on enzyme evolution. Genome Biol. 7:R39.
Wall DP, et al. 2005. Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci USA. 102:5483–5488.
Wheeler DL, et al. 2003. Database resources of the National Center for Biotechnology. Nucleic Acids Res. 31:28–33.
Wilke CO, Drummond DA. 2006. Population genetics of translational robustness. Genetics. 173:473–481.
Wilson AC, Carlson SS, White TJ. 1977. Biochemical evolution. Annu Rev Biochem. 46:573–639.
Wolf MY, Wolf YI, Koonin EV. 2008. Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution. Biol Direct. 3:40.
Wolf YI. 2006. Coping with the quantitative genomics 'elephant': the correlation between the gene dispensability and evolution rate. Trends Genet. 22:354–357.
Wolf YI, Carmel L, Koonin EV. 2006. Unifying measures of gene function and evolution. Proc Biol Sci. 273:1507–1515.
Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proc Natl Acad Sci USA. 106:7273–7280.
Zhou T, Drummond DA, Wilke CO. 2008. Contact density affects protein evolutionary rate from bacteria to animals. J Mol Evol. 66: 395–404.
Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence of proteins. In: Bryson V, Vogel HJ, editors. Evolving gene and proteins. New York: Academic Press. pp. 97–166.

Associate editor: William Martin