Highly Expressed and Slowly Evolving Proteins Share Compositional Properties with Thermophilic Proteins

Joshua L. Cherry*

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

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

Associate editor: Takashi Gojobori

Abstract

The sequences of proteins encoded by a genome evolve at different rates. A correlate of a protein's evolutionary rate is its expression level: highly expressed proteins tend to evolve slowly. Some explanations of rate variation and the correlation between rate and expression predict that more slowly evolving and more highly expressed proteins have more favorable equilibrium constants for folding. Proteins from thermophiles generally have more stable folds than proteins from mesophiles, and it is known that there are systematic differences in amino acid content between thermophilic and mesophilic proteins. I examined whether there are analogous correlations of amino acid frequencies with evolutionary rate and expression level within genomes. In most of the organisms analyzed, there is a striking tendency for more slowly evolving proteins to be more thermophile-like in their amino acid compositions when adjustments are made for variation in GC content. More highly expressed proteins also tend to be more thermophile-like by the same criteria. These results suggest that part of the evolutionary rate variation among proteins is due to variation in the strength of selection for stability of the folded state. They also suggest that increasing strength of this selective force with expression level plays a role in the correlation between evolutionary rate and expression level.

Key words: expression level, evolutionary rate, protein evolution, thermophilic.

Research article

Introduction

The forces that shape protein evolution are a central topic in molecular evolution. Within-genome differences in rates of protein evolution provide a window into these forces. Correlations between evolutionary rate and several other variables have been observed (Pál et al. 2001; Krylov et al. 2003; Rocha and Danchin 2004; Drummond et al. 2005; Drummond and Wilke 2008). Perhaps surprisingly, one of the best predictors of a protein's evolutionary rate is its expression level. Despite these correlational observations, the causes of rate differences remain uncertain.

Drummond et al. (2005) proposed that the negative correlation between evolutionary rate and expression level reflects selection against harmful effects of misfolded proteins. According to this hypothesis, the products of misfolding are toxic for reasons unrelated to the function of the protein. The products of errors in translation and transcription are particularly likely to misfold, because alteration of the amino acid sequence can make proper folding unfavorable. Selection against such toxic effects would be stronger for highly expressed proteins, simply because a given fraction of misfolding would correspond to a larger quantity of misfolded protein. This stronger selection would lead to both slower evolutionary rates and greater stability of the folded state for highly expressed proteins. These predictions were borne out by simulations of a model of this hypothesis (Drummond and Wilke 2008).

The more conventional view is that the main selective constraint on protein evolution is selection for function. Stability of the folded state is important to this kind of selection as well. Unfolded protein is obviously not functional (with the exception of some intrinsically disordered proteins [Dyson and Wright 2005]). Selection against sequence changes that largely abolish folding reduces the rate of evolution. Such changes are strongly deleterious and do not become fixed. However, more subtle differences in stability of the folded state can also be selectively important, and many of these will be weakly selected. For example, lowering the equilibrium constant for folding from 1,000 to 100 would, other things being equal, lead to nearly a 1% loss of protein function, with a selective cost that depends on functional aspects of the protein. This type of selection has been modeled by Chen and Shakhnovich (2009). Furthermore, unfolded protein is subject to degradation, so even a small fraction of unfolded protein at equilibrium might lead to a large decrease in the steady-state protein concentration, and hence a disproportionately large loss of protein function. Fast folding, slow unfolding, and stability to denaturing conditions are other, related targets of selection. Mistranslation might play a role in any of these types of selection. Stronger functional selection will lead to both slower evolution and greater stability. To the extent that functional constraint is stronger for highly expressed proteins, these will also tend to have greater stability.

Thus, several hypotheses predict that more slowly evolving and more highly expressed proteins tend to have more stable folded states. Such within-organism differences are reminiscent of the difference between mesophilic and thermophilic proteins, which also involves stability of the folded state. This is not to say that thermostability is precisely the same problem as greater stability at a particular temperature. As temperature increases, enthalpic changes of fixed size become less important for both rate constants and equilibrium constants. Furthermore, the thermodynamics of important interactions such as salt bridges and the hydrophobic effect exhibit complicated temperature dependence (Makhatadze and Privalov 1995; Elcock 1998). Nonetheless, direct empirical evidence confirms a connection between thermostability and increased stability at ordinary temperatures: even at ordinary temperatures, most thermophilic proteins have more stable folds (both kinetically and thermodynamically) than their mesophilic counterparts (Kumar and Nussinov 2001; Luke et al. 2007). Thus, sequence features that distinguish thermophilic from mesophilic proteins might also distinguish more stably folded mesophilic proteins from less stably folded mesophilic proteins. According to the evolutionary hypotheses discussed above, these features would also distinguish slowly evolving from rapidly evolving proteins and highly expressed from lowly expressed proteins.

The amino acid compositions of thermophilic proteins are systematically different from those of mesophilic proteins (Singer and Hickey 2003; Zeldovich et al. 2007). As argued by these authors, these differences likely reflect, at least in part, differences in the thermodynamics and kinetics of folding. I therefore investigated whether the same compositional differences are associated with differences in evolutionary rate and expression level within genomes. I computed the within-genome correlations between amino acid frequencies and evolutionary rate for several organisms, controlling for differences in GC content among genes. I did the same for correlations between amino acid frequencies and expression level. I compared these correlations to differences in amino acid frequencies between thermophilic and mesophilic proteins. For most of the organisms analyzed, there is substantial agreement between the directions of these correlations and the directions of thermophile/mesophile differences. This suggests that more slowly evolving and more highly expressed proteins do indeed have more stable folds.

Methods

Ortholog pairs for *Homo sapiens* and *Macaca mulatta*, *Drosophila melanogaster* and *Drosophila simulans*, and *Aspergillus fumigatus* and *Neosartorya fischeri*, along with expression data for human genes, were kindly provided by Yuri Wolf and are as described in Wolf et al. (2009). Ortholog pairs for *Escherichia coli* and *Salmonella typhimurium* were obtained from the Roundup server (Deluca et al. 2006), using default parameters (divergence = 0.2, Blast *E*-value = 1×10^{-20}). Ortholog pairs for *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* were as identified by Kellis et al. (2003) and described in the file at http://downloads.yeastgenome.org/sequence/fungal_genomes/S_paradoxus/other/MIT_paradoxus_hits.txt. *Caenorhabditis briggsae–Caenorhabditis remanei* ortholog information was downloaded from the TreeFam database (Ruan et al. 2008) and processed to eliminate pairs with bootstrap probability below 95% and to retain just one pair from each many:many ortholog relationship.

Saccharomyces cerevisiae mRNA levels were those reported by Holstege et al. (1998) and were downloaded from http://web.wi.mit.edu/young/pub/data/orf_transcriptome.txt. *Saccharomyces cerevisiae* protein abundance levels were from Ghaemmaghami et al. (2003). *Caenorhabditis elegans* mRNA data were from Hill et al. (2000), Supplemental table 2a, obtained from http://www.mcb.harvard.edu/hunter/Pubs/1053496_supplemental.zip. Expression levels reported for the eight developmental stages were averaged. *Caenorhabditis elegans* protein abundance data were from Schrimpf et al. (2009). *Escherichia coli* expression data were from Covert et al. (2004), Supplementary Data 7. Data for wild-type *E. coli* were used. The three aerobic data sets were averaged, as were the four anaerobic data sets. The resulting aerobic and anaerobic means were then averaged to produce the values used in this study. *Drosophila melanogaster* mRNA data were obtained from FlyAtlas (Chintapalli et al. 2007). Expression levels for whole fly were used. For probes that corresponded to multiple splicing variants, one was chosen at random. *Drosophila melanogaster* protein abundances were from Brunner et al. (2007).

Analyses were performed with the aid of the Python programming language along with NumPy (Oliphant 2007; http://numpy.scipy.org) and the Python interface to the *National Center for Biotechnology Information* (NCBI) C++ Toolkit. Sequences were obtained either from NCBI databases, files provided by third parties, or a third-party database. Genes with low-complexity protein sequences (entropy less than 2.5 nats) were excluded from the analysis. Except where noted, a limit on MaxH (Boyd et al. 1998) of 1.4 was imposed in an effort to exclude most membrane proteins. Protein sequences of apparent orthologs were aligned with MUSCLE (Edgar 2004) using the default settings, and these alignments were propagated back to the coding sequences. Pairs whose alignments indicated frame-changing differences or otherwise variable quality were eliminated from consideration. Protein distances were calculated using the protdist program of the PHYLIP package (Felsenstein 2005), using a gamma distribution of rates with a fixed shape parameter of 1. dN and dS were estimated using the CODEML program of the PAML package (Yang 1997) with constant rates and CodonFreq = 2. For correlations with rates of evolution, genes with $d\text{S} > 2$ were excluded, except that $d\text{S} > 4$ was the cut-off for *C. briggsae–C. remanei* pairs because of the greater divergence between these organisms.

Correlation results controlled for GC content were obtained by computing Spearman's rank-order correlation coefficient for the residuals of third-degree polynomial fits

of the variables to GC fraction. Results were found to depend only weakly on the degree of the polynomial.

Kernel smoothing regression (fig. 1) was carried out in Octave (http://www.octave.org). Vertical values of each plot were calculated as weighted averages of the vertical values of the data points (each data point representing one protein), with the weights given by a Gaussian function centered at the corresponding horizontal value. The standard deviation (SD) of the Gaussian was 1/ln(10), which corresponds to a standard deviation of one on a natural log scale.

Results

General Approach and Detailed Example

Singer and Hickey (2003) compared the amino acid compositions of proteins from several mesophilic and thermophilic species. Differences were individually statistically significant at the 5% level for 11 of the 20 amino acids: E, I, K, V, and Y were overrepresented in thermophiles (compared with mesophiles) and A, C, H, Q, T, and W were underrepresented. I will refer to this set of amino acids, together with their classifications as overrepresented or underrepresented in thermophiles, as SH-11. This grouping of the amino acids is apparently unrelated to their metabolic costs, which might influence correlations between amino acid frequencies and expression level (Akashi and Gojobori 2002). Similar results follow from the correlations with optimal growth temperature reported by Zeldovich et al. (2007, their Table S5a); if only statistically significant correlations are considered (ignoring the lack of phylogenetic independence), the only differences are that S and D have negative correlations with optimal growth temperature and W has no significant correlation. The results presented here are based on those of Singer and Hickey (2003) (SH-11). Use of the results of Zeldovich et al. (2007) instead, or use of the union or intersection of the two sets of amino acids, would strengthen some of the results and weaken others, but leave the overall conclusions unchanged.

The analyses presented here are based on rank-order correlation coefficients between the frequency of each amino acid and the variable of interest (a measure of evolutionary rate or expression level), adjusted for GC content. Only those amino acids for which this correlation and the thermophile/mesophile difference were both statistically significant at the 5% level were considered. The total number of such amino acids and the number for which the correlation had the expected direction were tabulated. The "expected direction" means that slowly evolving or highly expressed proteins are more like thermophilic proteins. For example, for tyrosine (Y), which is overrepresented in thermophiles, the expectation is a negative correlation with protein evolutionary rate and a positive correlation with expression level.

The details of one such analysis are shown in table 1. For human proteins, correlation coefficients (controlled for GC content, as explained below) between each amino acid frequency and a measure of evolutionary rate are shown, along with their *P* values. The measure of evolutionary rate was a protein distance calculated for each human protein and an apparent *M. mulatta* ortholog. Of the 11 amino acids whose frequencies differ significantly between thermophiles and mesophiles, 8 (all but E, K, and H) had statistically significant correlations ($P < 0.05$) with this measure of evolutionary rate. Of these eight correlations, all but one had the predicted sign, as indicated in the table. This much agreement is unlikely to occur by chance: for a one-tailed binomial test of the null hypothesis that agreement and disagreement are equally likely, the *P* value for this result (7/8 matching the expectation) is 0.035.

Table 1. Correlation Results for Human Protein Distances, Controlling for GC Content.

Amino acid	Correlation coefficient	*P* value	Agreement with prediction
A	−0.039	0.00027	Disagrees
C	0.107	2.9×10^{-23}	Agrees
D	−0.149	1.7×10^{-43}	
E	−0.013	0.24	ns
F	0.003	0.75	
G	−0.031	0.0039	
H	0.008	0.46	ns
I	−0.111	1.6×10^{-24}	Agrees
K	−0.011	0.33	ns
L	0.053	9.5×10^{-07}	
M	−0.087	9.8×10^{-16}	
N	−0.095	2.5×10^{-18}	
P	0.052	1.6×10^{-06}	
Q	0.057	1.2×10^{-07}	Agrees
R	0.077	1.3×10^{-12}	
S	0.016	0.13	
T	0.033	0.0027	Agrees
V	−0.069	1.6×10^{-10}	Agrees
W	0.127	5.8×10^{-32}	Agrees
Y	−0.111	7.5×10^{-25}	Agrees

NOTE.—ns, not significant.

The relationships between mean amino acid frequencies and evolutionary rate are illustrated by figure 1. These curves, which were produced by kernel smoothing regression (see Methods), may be compared with the correlation results in table 1. Figure 1*A* shows a generally downward trend as estimated evolutionary rate increases for three of the amino acids that are more frequent in thermophilic proteins (I, V, and Y). For E and K there is not a clear trend. Visual inspection might suggest an overall downward trend, but the estimated correlation coefficients, though negative, are not statistically significant. In figure 1*B*, an increasing trend is obvious for three of the amino acids that are less frequent in thermophilic proteins: C, T, and W. The curve for H is fairly flat and peaks near the middle, which is consistent with its low and nonsignificant correlation. The curve for A is strikingly nonmonotonic (increasing in some places and decreasing in others). It is decreasing in the middle of the range of evolutionary rate, where the bulk of the data points lie, which is consistent with the discordant negative correlation for A. The curve for Q also displays marked nonmonotonicity, but is rising in the middle of the data range, which is consistent with the positive correlation for Q. Figure 1*C* shows the relationships for the other amino acids, which can also be reconciled with the correlation results in table 1.

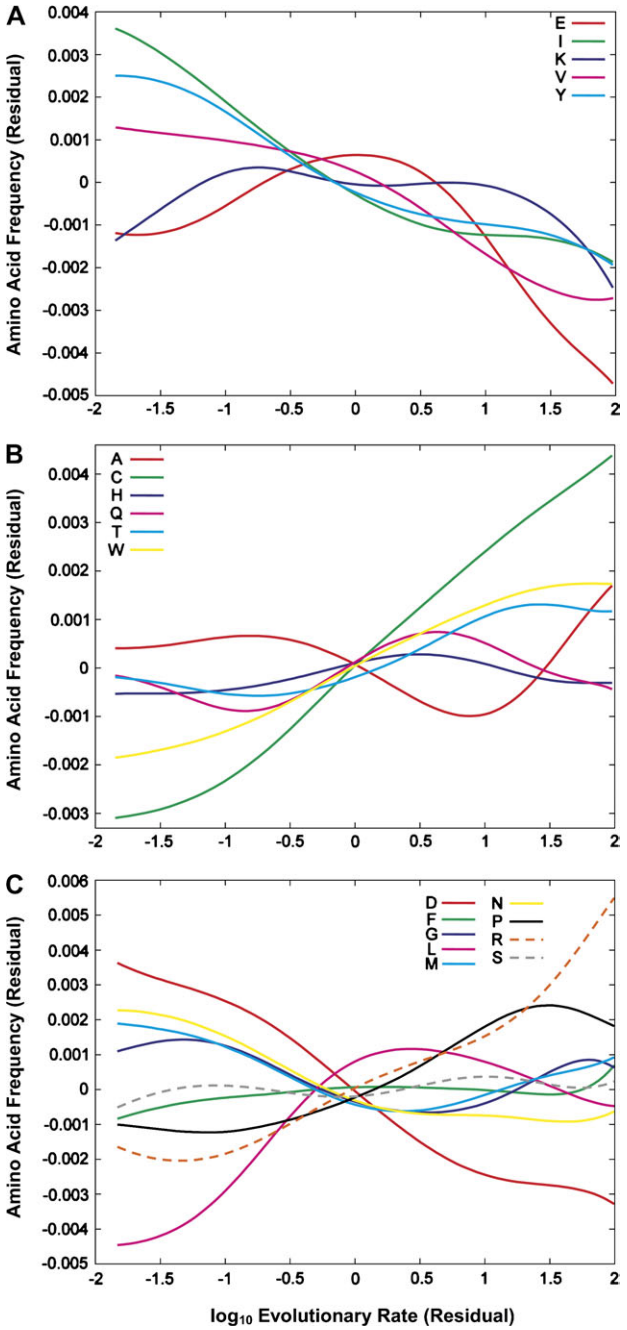

Fig. 1. The relationship between amino acid frequency and evolutionary rate among human proteins. The curve for each amino acid conveys how its frequency varies with protein sequence distance when GC content is taken into account. Each curve was produced by smoothing the data with a Gaussian kernel with an SD of 1/ln(10). The raw data points were the residuals of cubic polynomial fits of amino acid frequencies and the logarithm (base 10) of sequence distance to GC fraction. (*A*) Amino acids that are overrepresented in thermophiles. (*B*) Amino acids that are underrepresented in thermophiles. (*C*) Other amino acids.

Amino Acid Frequencies and Evolutionary Rate

Table 2 summarizes the results of analysis of correlations between amino acid frequencies and evolutionary rates for six organisms. Each result summarizes an analysis of the type shown in table 1. Results are shown for three measures of protein evolutionary rate (protein distance [calculated from protein alignments, without consideration of nucleotide sequences], *d*N, and *d*N/*d*S). For comparison, results for a measure of synonymous rate (*d*S) are also shown.

Table 2. Correlation Results for Amino Acid Frequencies and Evolutionary Rate.

		Protein evolutionary rate			
	Number of genes	Protein distance	*d*N	*d*N/*d*S	*d*S
Home sapiens	8,502	7/8	7/8	9/9	3/9
Drosophila melanogaster	5,369	6/7	6/7	7/8	2/11
Aspergillus fumigatus	5,532	10/11	10/10	11/11	6/11
Saccharomyces cerevisiae	3,367	7/8	6/7	6/7	3/5
Escherichia coli	1,720	9/9	9/9	8/8	9/10
Caenorhabditis briggsae	4,922	5/10	5/10	6/8	3/9
C. briggsae, *d*S < 1	878	7/7	7/7	5/5	2/2

Note.—For each organism and each rate measure, the number of statistically significant correlations that have the predicted sign is given by the numerator and the total number of significant correlations for which a prediction exists is given by the denominator.

For all of the organisms other than *C. briggsae*, there is overwhelming agreement between the observed and expected directions of correlation for all three measures of protein evolutionary rate. In no case does more than one amino acid have a significant correlation in the "wrong" direction, whereas in all cases at least six amino acids, and in one case all 11, have correlations in the expected direction. Except in the cases with 6/7 agreements, each of these results is individually statistically significant according to a one-tailed binomial test, with *P* values ranging from 0.035 (for 7/8 agreements) to 0.00049 (for 11/11 agreements). The cases with 6/7 agreements approach statistical significance ($P = 0.063$). Thus, the compositional differences that distinguish thermophilic from mesophilic proteins tend also to distinguish more slowly evolving proteins from more rapidly evolving proteins.

For *C. briggsae*, no strong trend toward agreement is apparent for the full set of *C. briggsae*–*C. remanei* apparent orthologs. This may reflect the fact that *C. briggsae* and *C. remanei* are more diverged than the other pairs of organisms used in this study (*C. elegans* and *C. briggsae* are even more highly diverged, and comparing them yields similar results). If the analysis is restricted to ortholog pairs with estimated $dS < 1$, a strong trend toward agreement is observed (table 2), comparable to that found in other organisms and statistically significant ($P = 0.031$ for *d*N/*d*S and $P = 0.0078$ for the other measures of protein evolutionary rate). This result is difficult to interpret, but it suggests that whatever force is responsible for the effect observed in the other organisms also operates to some extent in *Caenorhabditis*.

The correlations used in these analyses are controlled for the GC content of the coding sequences (see Methods for details). Differences in GC content among genes in the same organism can lead to corresponding differences in amino acid frequencies, and evolutionary rates correlate negatively or positively with GC content, depending on

	Amino Acid	*H. sapiens*	*D. melanogaster*	*A. fumigatus*	*S. cerevisiae*	*E. coli*	*C. briggsae*, dS<1	*C. briggsae*
More Frequent in Thermophiles	E	ns	ns	-	ns	-	-	+
	I	-	-	-	-	-	-	-
	K	ns	-	-	ns	-	-	ns
	V	-	-	-	-	-	-	-
	Y	-	-	-	-	ns	-	-
Less Frequent in Thermophiles	A	-	ns	ns	-	+	ns	-
	C	+	+	+	ns	+	ns	-
	H	ns	ns	+	+	+	ns	-
	Q	+	+	+	+	+	+	+
	T	+	ns	+	+	ns	+	+
	W	+	-	+	ns	+	ns	-
Others	D	-	-	ns	ns	-	-	-
	F	ns	-	-	-	ns	ns	ns
	G	-	-	-	-	-	ns	-
	L	+	-	ns	-	+	-	-
	M	-	-	-	-	-	-	ns
	N	-	-	-	+	ns	ns	+
	P	+	+	+	+	+	+	+
	R	+	+	+	+	ns	-	ns
	S	+	+	+	+	+	+	+

Fig. 2. Summary of correlation results for *d*N across organisms. For each correlation, "+" indicates a statistically significant ($P < 0.05$) positive correlation, "−" indicates a significant negative correlation, and "ns" indicates that the correlation was not statistically significant. Negative correlations, which are predicted for amino acids more common are thermophiles, are also indicated by orange coloring, and positive correlations are indicated by blue coloring.

the organism. Thus, it is necessary to control for GC content. Table S1 (Supplementary Material online) summarizes the results of using correlations that are not controlled for GC content. For *D. melanogaster* and *S. cerevisiae*, no strong effect is apparent; controlling for GC content exposes an otherwise hidden phenomenon. For *H. sapiens*, *A. fumigatus*, *E. coli*, and *C. briggsae* ($dS < 1$), the effect is mostly observable even without correction for GC content. In these cases, controlling for GC content provides assurance that the observed effect is not a simple artifact of compositional variation. As an additional check, I performed the analysis for *H. sapiens* controlling for the GC content of introns in the coding sequence, restricting the analysis to the 7576 genes with at least 1,000 bp of intron sequence. The results again show a strong effect: 6/7, 7/8, and 7/8 of the significant correlations go in the expected direction for protein distance, *d*N, and *d*N/*d*S, respectively.

Because systematic differences between membrane and nonmembrane proteins might affect the results, likely membrane proteins were excluded from the analysis. This exclusion was based on MaxH, a simple metric devised by Boyd et al. (1998) that distinguishes between the two types of protein on the basis of sequence. The correlation analysis was restricted to proteins with MaxH < 1.4, which should eliminate the vast majority of membrane proteins (along with many nonmembrane proteins). Results for the full sets of proteins, not filtered by MaxH, are shown in supplementary table S2 (Supplementary Material online). For *H. sapiens*, *A. fumigatus*, and *E. coli*, all of these results are identical to or stronger than the corresponding results for the restricted set of proteins (table 2). However, the results for *D. melanogaster*, *S. cerevisiae*, and *C. briggsae* ($dS < 1$) are significantly weakened. It may be significant that *D. melanogaster* and *S. cerevisiae* are also the only organisms that required correction for GC content in order to show the observed effect.

The correlations of *d*N with each amino acid frequency are summarized graphically in figure 2. The abundance of negative correlations for amino acids that are more frequent in thermophilic proteins, and positive correlations for amino acids with the opposite tendency, is apparent. Neglecting *C. briggsae* (but including the *C. briggsae* $dS < 1$ results), the only amino acids whose frequencies ever correlate significantly in the "wrong" direction are alanine (A) and tryptophan (W), the latter of which had no significant correlation with optimal growth temperature according to Zeldovich et al. (2007). Despite being discordant in some organisms, both A and W correlate in the expected direction in other organisms. For each of the other nine amino acids in SH-11, there is a consensus of sorts: all of the significant correlations have the same sign, and there is more than one significant correlation. In all nine cases in which there is a consensus, the consensus is in accord with the hypothesis that more slowly evolving proteins are more like thermophilic proteins.

Trends in the direction of correlation are also apparent for some amino acids outside of SH-11. Most notably, P and S consistently correlate positively with *d*N, and M consistently correlates negatively (even if initial methionines are excluded from the analysis). The positive correlations for S are in accord with the significant negative correlation with optimal growth temperature reported by Zeldovich et al. (2007). However, the correlations for D, the other amino acid specific to the Zeldovich et al. (2007) set, are discordant in the cases where they are statistically significant.

As mentioned above, the SH-11 classifications of the amino acids bear no clear relationship to their metabolic costs as calculated by Akashi and Gojobori (2002). We may ask whether, considering all the amino acids, the observed correlations tend to reflect metabolic costs, perhaps due to the correlation of evolutionary rate with expression level. There is no significant difference between the costs of the nine amino acids for which there is a consensus negative correlation and the seven for which there is a consensus positive correlation ($P = 0.22$, Mann–Whitney *U*-test). Thus, the correlations do not appear to reflect the metabolic costs of amino acids.

The picture is much the same for the correlations with protein distance and *d*N/*d*S (Supplementary figs S1 and S2, Supplementary Material online). The most notable difference is that threonine, rather than tryptophan, is discordant for protein distance in *D. melanogaster*. Alanine

Table 3. Correlation Results for Amino Acid Frequencies and Expression Level.

	Number of genes	Correlation results	*P* value	Discordant amino acid(s)
Homo sapiens (EST counts)	8,143	6/7	0.063	A
H. sapiens (microarray)	7,752	9/10	0.011	A
Drosophila melanogaster (mRNA)	7,791	8/9	0.020	A
D. melanogaster (protein)	5,450	9/10	0.011	A
Saccharomyces cerevisiae (mRNA)	3,542	8/10	0.055	A, Y
S. cerevisiae (protein)	2,820	8/9	0.020	A
Escherichia coli	2,712	8/9	0.020	Y
Caenorhabditis elegans (mRNA)	7,053	5/6	0.109	A
C. elegans (protein)	6,269	9/10	0.011	A

NOTE.—The number of statistically significant correlations that have the predicted sign is shown, along with the total number of significant correlations for which a prediction exists. Discordant amino acids are those with significant correlations in the opposite of the predicted direction.

remains the only amino acid to be discordant in any other organism.

Amino Acid Frequencies and Expression Level

Table 3 and Supplementary figure S3 (Supplementary Material online) summarize the results of correlation of amino acid frequencies with expression levels, again controlling for GC content. Just as for evolutionary rate, there is strong agreement between the observed directions of the correlations and the directions predicted on the basis of the differences between thermophiles and mesophiles reported by Singer and Hickey (2003). In all but one case, just one amino acid was discordant. Six of the nine results are individually statistically significant. Some of the remaining three come close, and in each of these three cases a different measure of expression level in the same organism yields a statistically significant result.

In every analysis summarized in table 3, the frequency of either alanine or tyrosine correlates in the opposite of the predicted direction. The recurring discordance of alanine is reminiscent of the results for protein evolutionary rate. However, in the case of expression level, alanine never correlates in the predicted direction: its correlation is either discordant or not statistically significant. There is, then, a consensus for alanine (in the sense defined in the previous section), and this consensus is discordant. For tyrosine, in contrast, there is no consensus among the results for expression level. As is evident from Supplementary figure S3 (Supplementary Material online), there is consensus for ten of the SH-11 amino acids, and nine of these are in accord with the hypothesis that the compositions of highly expressed proteins tend to be similar to those of thermophilic proteins.

Stronger selection against use of metabolically costly amino acids in highly expressed proteins might affect these correlations. As Supplementary figure S3 (Supplementary Material online) shows, there are consensus positive correlations for seven amino acids and consensus negative correlations for eight. There is no statistically significant difference between the costs of these two sets of amino acids ($P = 0.15$, Mann–Whitney *U*-test) according to the values calculated by Akashi and Gojobori (2002). Selection for low-cost amino acids does not appear to explain the correlations.

Discussion

Within the genomes analyzed, the amino acid composition of a protein correlates with its evolutionary rate and expression level. For most of these genomes, the correlations, controlling for GC content, tend to mirror the compositional differences between mesophilic and thermophilic proteins. The frequencies of amino acids that are overrepresented in thermophiles tend to correlate negatively with evolutionary rate and positively with expression level. For amino acids that are rarer in thermophilic proteins, the correlations tend to go in the opposite direction. Thus, both highly expressed proteins and slowly evolving proteins tend to be more like thermophilic proteins in their amino acid compositions.

Although other interpretations are possible, these results strongly suggest that more slowly evolving and more highly expressed proteins tend to have more stable folded states (i.e., more favorable equilibrium constants for folding). This suggests that evolutionary rate is determined in part by the strength of selection for folding stability and that the reason for the observed negative correlation between expression level and evolutionary rate is that higher expression leads to stronger selection for stability. In slight variations of this interpretation, the target of selection is not thermodynamic stability per se, but a related attribute such as high speed of folding, low rate of unfolding, or rigidity of the folded structure.

A variety of hypotheses would explain stronger selection for proper folding of more highly expressed proteins. Drummond et al. (2005) proposed that toxic effects of the misfolded products of mistranslation were the dominant force. Selection against loss of protein function would also explain the effect, provided that the loss of a given fraction of protein functionality tends to have a greater cost for more highly expressed proteins, as has been proposed (Rocha and Danchin 2004). Whether the important cost of misfolding is toxicity or loss of protein function, mistranslation might or might not be important.

Although the results presented here do not distinguish among hypotheses that invoke selection for proper protein folding, they do support this class of hypotheses against alternatives. For example, the hypothesis that the correlation between expression level and evolutionary rate is due to selection for translational efficiency (Akashi 2001) does not predict the observed effect (although it is not strictly incompatible with it). Furthermore, properties unrelated to free energy of folding are important for protein function. Any or all of these might be important for explaining rate differences among proteins and the correlation between rate and expression level. The results presented here, however, point specifically at more favorable folding or a related

property as an important factor. The same can be said of selection against toxic effects of proteins, which in principle could act on something unrelated to stable folding.

Deciding among explanations for the correlation between expression level and evolutionary rate will require further evidence. The results presented here suggest that the correct explanation will involve selection on some aspect of protein folding.

Supplementary Material

Supplementary tables S1 and S2 and figures S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

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

References

Akashi H. 2001. Gene expression and molecular evolution. *Curr Opin Genet Develop*. 11:660–666.

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

Boyd D, Schierle C, Beckwith J. 1998. How many membrane proteins are there? *Protein Sci*. 7:201–205.

Brunner E, Ahrens CH, Mohanty S, Baetschmann H, Loevenich S, et al. 2007. A high-quality catalog of the *Drosophila melanogaster* proteome. *Nat Biotech*. 25:576–583.

Chen P, Shakhnovich EI. 2009. Lethal mutagenesis in viruses and bacteria. *Genetics* 183:639–650.

Chintapalli VR, Wang J, Dow JA. 2007. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. *Nat Genet*. 39:715–720.

Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. 2004. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429:92–96.

Deluca TF, Wu IH, Pu J, Monaghan T, Peshkin L, Singh S, Wall DP. 2006. Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics* 22:2044–2046.

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

Dyson HJ, Wright PE. 2005. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*. 6:197–208.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792–1797.

Elcock AH. 1998. The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins. *J Mol Biol*. 284:489–502.

Felsenstein J. 2005. *PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author*. Seattle: Department of Genome Sciences, University of Washington.

Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS. 2003. Global analysis of protein expression in yeast. *Nature* 425:737–741.

Hill AA, Hunter CP, Tsung BT, Tucker-Kellogg G, Brown EL. 2000. Genomic analysis of gene expression in *C. elegans*. *Science* 290:809–812.

Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95:717–728.

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

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.

Kumar S, Nussinov R. 2001. How do thermophilic proteins deal with heat? *Cell Mol Life Sci*. 58:1216–1233.

Luke KA, Higgins CL, Wittung-Stafshede P. 2007. Thermodynamic stability and folding of proteins from hyperthermophilic organisms. *FEBS J*. 274:4023–4033.

Makhatadze GI, Privalov PL. 1995. Energetics of protein structure. *Adv Protein Chem*. 47:307–425.

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.

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

Ruan J, Li H, Chen Z, Coghlan A, Coin LJ, et al. 2008. TreeFam: 2008 update. *Nucleic Acids Res*. 36:D735–D740.

Schrimpf SP, Weiss M, Reiter L, et al. (13 co-authors). 2009. Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol*. 7:e48.

Singer GA, Hickey DA. 2003. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* 317:39–47.

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.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*. 13:555–556.

Zeldovich KB, Berezovsky IN, Shakhnovich EI. 2007. Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput Biol*. 3:e5.