Why Does a Protein's Evolutionary Rate Vary over Time?

Xiangjun Du, David J. Lipman, and 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: February 7, 2013

Abstract

The sequences of different proteins evolve at different rates. The relative evolutionary rate (ER) of a single protein also changes over evolutionary time. The cause of this ER fluctuation remains uncertain, and study of this phenomenon may shed light on protein evolution more broadly. We have characterized ER fluctuation in mammals and *Drosophila*. We found little correlation between the amount of rate variation observed for a protein and such factors as its expression level or phylogenetic distribution. Perhaps more surprisingly, we found little correlation between our measure of rate variation and ER itself. We also investigated the extent to which the ERs of different domains of a protein vary independently. We found that rates of different domains do tend to vary together. In fact, rates at positions in different domains are coupled just as strongly as rates at equally distant positions in the same domain. These findings provide clues to the protein evolutionary process.

Key words: protein evolution, evolutionary rate, overdispersion, molecular clock.

Introduction

The role of various types of selection in protein sequence evolution is a matter of ongoing debate. The "neutralist/selectionist controversy" concerns whether adaptation drives sequence evolution (selectionism) or selection serves mainly as a constraint on sequence evolution (neutralism) (Kreitman 1996; Ohta 1996). Within the neutralist framework, the nature of selective constraint is not settled.

There is considerable variation in the evolutionary rates (ERs) of proteins encoded by the same genome. This variation is commonly interpreted as an indication of differences in selective constraints. Perhaps surprisingly, measures of a protein's contributions to fitness do not correlate well with its ER. The best correlate of a protein's ER is its expression level; highly expressed proteins tend to evolve slowly (Pál et al. 2001; Krylov et al. 2003; Rocha and Danchin 2004; Drummond et al. 2006; Drummond and Wilke 2008).

Temporal variation in the ERs of individual proteins provides another window into selection on protein sequence. The evolution of many protein sequences is overdispersed: The number of sequence changes occurring along a lineage is more variable than expected with a constant rate of independent substitutions, which predicts a Poisson distribution of the number of substitutions, and hence equality of the variance and the mean (index of dispersion equal to 1) (Ohta and Kimura 1971; Langley and Fitch 1974). This phenomenon remains when genome-wide, lineage-specific changes in rate are accounted for (Gillespie 1989). Some have taken the existence and characteristics of overdispersion to be evidence that protein sequence evolution is largely driven by adaptation (Gillespie 1989, 1991). Others have attempted to explain overdispersion in terms of nearly neutral evolution (Takahata 1987; Cherry 1998; Bloom et al. 2007).

Rate variation might result from changes in the protein sequence itself or from external factors. The "Fluctuating Neutral Space" model of Takahata (1987) posits that nearly neutral changes to the protein sequence alter the subsequent rate of nearly neutral substitution, leading to overdispersion. A more specific hypothesis (Bloom et al. 2007) attributes changes in ER to changes in the stability of the protein's folded state: Stabilizing changes to the sequence lead to greater tolerance of subsequent destabilizing changes, and hence increase ER, and destabilizing changes have the opposite effect. Alternatively, extrinsic factors such as changing environment, changing lifestyle, or genetic changes elsewhere in the genome, all of which may alter the selective forces acting on a protein sequence, might be responsible for overdispersion. Bedford et al. (2008) analyzed ER variation in mammals, flies, and yeast, and concluded that it is driven in large part by changes to the protein sequence itself.

Here, we compare the ERs of orthologs in different parts of the phylogenetic tree, examining whether various factors correlate with the tendency of a protein's ER to vary. We find that several factors that are correlated with ER are not strongly

Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution 2013.

correlated with variation in ER, that variation in ER does not correlate strongly with ER itself, and that ERs of distinct domains of the same protein tend to vary together over time. These findings are evidence against the hypothesis that changes in a protein's ER are primarily due to changes in the protein's sequence, especially the hypothesis that ER changes are the result of changes to the stability of the protein's folded state.

Materials and Methods

Genomic Data

We gathered protein-coding sequences from four mammals and four *Drosophila* species. The mammalian species were *Homo sapiens* (Homsa), the rhesus macaque *Macaca mulatta* (Macmu), *Mus musculus* (Musmu), and *Rattus norvegicus* (Ratno). We refer to Homsa and Macmu, and the phylogenetic branches that connect them, as "primate." Similarly, Musmu and Ratno are referred to as "rodent." The *Drosophila* species were *D. melanogaster* (Drome), *D. simulans* (Drosi), *D. yakuba* (Droya), and *D. erecta* (Droer). We refer to Drome and Drosi, which form a clade, as "fly1," and refer to Droya and Droer, which also form a clade, as "fly2."

For all but two of the species (Macmu and Drosi), we obtained predicted proteins from the National Center for Biotechnology Information (NCBI) RefSeq database (Pruitt et al. 2007; Sayers et al. 2011). For each protein, we obtained the protein sequence and the corresponding coding sequence. We also determined the chromosomal locations of the genes and, for the mammals, their intron/exon structures.

For *D. simulans* (Drosi), we used coding sequences assembled from RNA-seq data using the Drome coding sequences as templates. Paired RNA-seq reads were obtained from the SRA database, project SRP007818. These Drosi reads were aligned to a set of Drome coding sequences using BLAST with command-line parameters "-word_size 16 -best_hit_overhang 0.1 -best_hit_score_edge 0.1." Alignments with length less than 60 bp were discarded. Mate pairs with correct orientation on the transcript and unique read placement were retained. A sequence was constructed for each transcript using a consensus approach. Where possible, gaps in the sequence were filled with the aid of unused reads.

For Macmu, we used the genome sequence of the Chinese rhesus (Yan et al. 2011), which has higher sequencing coverage than the Indian rhesus sequence (Rhesus Macaque Genome Sequencing and Analysis Consortium et al. 2007). Coding sequences were identified on the basis of alignments with the high-confidence human mRNA sequences in the RefSeq database (accessions beginning with "NM"). These were aligned to the Chinese rhesus genome using Splign (Kapustin et al. 2008). The best placement of each mRNA sequence was projected onto the rhesus genome, and the resulting rhesus coding sequence was retained if it was consistent with coding sequence of the human mRNA, that is, if it constituted an open reading frame with start and stop codons at the projected positions. In addition, we used a few gene models from the original annotations of this genome, which included some genes that were not found by our analysis and yet were acceptable by our filtering criteria (discussed later).

Orthologs and Alignment

For each pair of mammalian or fly genomes, reciprocal BLASTP searches (Altschul et al. 1997) were performed on proteins with *E* value threshold of 1e−06, and putative orthologs were identified as bidirectional best hits (Tatusov et al. 1997). Sets of four orthologs were identified by the requirement that each protein was the reciprocal best hit of each of the three other proteins in the set (i.e., each of the six possible pairings corresponded to a reciprocal best hit relationship).

Orthologous protein sequences were aligned using the MUSCLE program (Edgar 2004). The protein sequence alignments were then used to construct corresponding alignments of the coding sequences.

We filtered the mammalian alignments based on the exon structures of the coding sequences. This filtering consisted of two components. First, we required that amino acids encoded by the same exon in one organism were not aligned to amino acids encoded by different exons in another. Second, we required at least 50% amino acid identity for the aligned positions of any exon. These criteria were applied pairwise to the human/mouse, human/macaque, and mouse/rat components of every alignment.

To avoid undue influence of families of paralogs, we chose subsets of genes such that no pairs of paralogs with greater than 60% amino acid identity were included. Identity higher than 60% within any of the four genomes was sufficient to prohibit inclusion of a pair of ortholog groups. Paralogs were identified by within-genome all-against-all BLASTP searches. A pair of paralogs usually produced two hits, with query and subject interchanged; the larger of the two values of percent identity was used in these cases. We found that using a more conservative cutoff of 50% identity, or liberally including all genes without regard to presence of paralogs, produced substantially the same results.

Evolutionary Rate

ER parameters d*N*/d*S*, d*N*, and d*S* were estimated using CODEML program from the PAML package (Yang 1997), with separate d*N*/d*S* ratios for each branch and CodonFreq = 2. This was done for each ortholog set based on its corresponding CDS alignments. For most of the analyses, the calculations were based on the phylogenetic tree of the mammals or flies (fig. 1), yielding estimates for each of the five branches of the tree. Then, d*N* for each lineage (primate, rodent, fly1, or fly2) was calculated as the sum of d*N* values

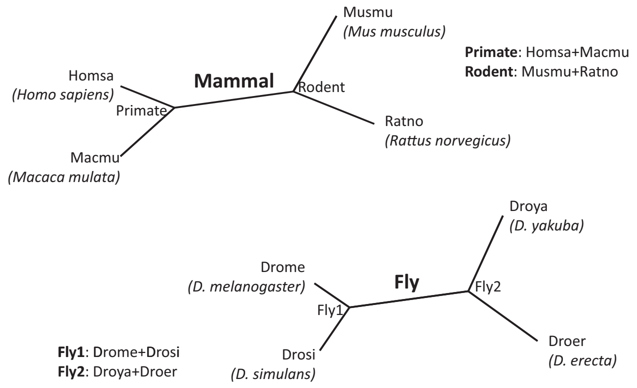

Fig. 1.—Phylogenetic relationships among the mammalian and *Drosophila* species used in this study. For both trees, the root lies along the internal branch.

from the two terminal branches of that lineage. Values of d*S* were obtained analogously. The d*N*/d*S* for each lineage was taken to be the ratio of the d*N* and d*S* values obtained in this way. For the domain analysis, d*N*/d*S*, d*N*, and d*S* values were instead calculated separately for each pair of closely related species. Genes with d*N*/d*S*, d*N*, or d*S* values more than three standard deviations from the mean were excluded from the analyses.

Our measure of temporal variation of ER was based on the ratio of the estimated ERs in two lineages (primate and rodent, or fly1 and fly2). If relative ERs were unchanging and there were no errors in the estimates, this ratio would be the same for all genes (for d*N* and d*S*, it would correspond to a ratio of branch lengths). As a measure of rate variability, we took the absolute deviation of the logarithm of this ratio from an expected value. The expected value was simply the logarithm of the geometric mean ratio (equivalently, the mean of the logarithm of the ratio) for most of the analyses. For the correlations between rate and rate variability, we used the absolute values of the deviations from a regression line in the logarithmic domain. This analysis is similar to the Breusch–Pagan test, but it involves a rank-order correlation rather than Pearson's product-moment correlation.

Expression and Age Class of Genes

Expressed sequence tag (EST) counts for Homsa and Musmu were downloaded from the NCBI Unigene database (Sayers et al. 2012). The mRNA data for Drome were downloaded from FlyAtlas (Chintapalli et al. 2007) and the mean values were used. The protein abundance data were obtained from the work of Schrimpf et al. (2009). The age class of genes from Homsa and Drome were kindly provided by Y. Wolf and described in the work of Wolf et al. (2009). The age classes were numbered, with 1 corresponding to the least phylogenetic depth and 7 to the greatest.

Protein Domain Analysis

Domain architecture was determined using the Conserved Domain Database based on the superfamily classification (Marchler-Bauer et al. 2011) for each protein. Partial domain matches were not used. Only domains that were detected in all the mammal or fly species were used, and the extent of the domain was taken to be the intersection of regions identified in the four species.

To gauge the effects of chromosomal proximity, we performed a comparable analysis with pairs of nearby genes. Adjacency of gene pairs in mammals was assessed solely by their position in the Homsa genome. A fraction of pairs categorized as adjacent in this way will not be nearby in one or more of the other mammals due to genome rearrangements. Similarly, adjacency in flies was assessed using the Drome genome sequence.

To compare within-domain to between-domain effects, we constructed sets of three regions from a coding sequence, two of which encoded regions of one protein domain and the third of which encoded a region in an adjacent domain. These regions were chosen such that all three had the same length and the distance along the protein sequence from the first to the second was the same as the distance from the second to the third. Only one set of regions was constructed for each ortholog set. Regions were chosen so as to maximize their length.

Statistical Analysis

Spearman's rank-order correlation was used for correlation analysis. The statistical significance of a correlation was assessed by a permutation test. The values for one variable were randomly permuted, so that each value for one variable was randomly paired with a value for the other variable. The correlation for the randomized data was then calculated. This process was repeated 10,000 times. The *P* value was taken to be the fraction of these correlations that had an absolute value at least as large as that of the correlation of the actual (unpermuted) data. For the correlations of ER with rate variability, the regression line was calculated separately for every permutation, and, in the sampling analysis (discussed later), for every combination of a permutation and a random sample.

As discussed in Results, genes that evolve more slowly will exhibit more sampling variance for rate estimates, leading to greater deviations of the observed ratio of rates from the central value. This may make a variable appear to correlate with rate variability when it merely correlates with rate. To eliminate this effect, we equalized sampling variance by sampling some fixed total number, *n*, of amino acid changes from each gene with at least *n* changes in the terminal branches. These were drawn at random, without replacement, from the changes in the two lineages (e.g., primate and rodent), "thinning" the observed substitutions so that the same total number remained for all genes. Equivalently, the number of the *n* changes assigned to one lineage was drawn from a

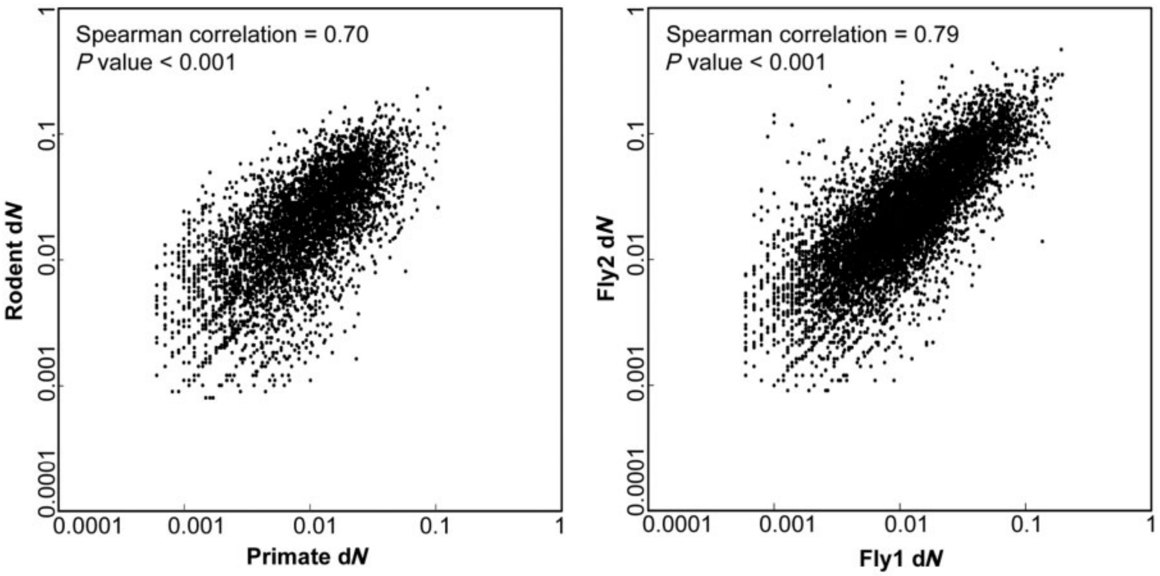

Fig. 2.—Relationships between protein ERs in different lineages. Each point represents a set of orthologous genes.

hypergeometric distribution, with the remainder assigned to the other. Suppose, for example, that $n = 10$. Consider a mammalian ortholog set with 7 inferred amino acid changes in the primate lineage and 15 in the rodent lineage. After sampling, the number of changes in the primate lineage will be an integer between 0 and 7, inclusive, with probabilities given by a hypergeometric distribution. The number in the rodent lineage will correspondingly range from 10 to 3, so that the sum for the two lineages is necessarily 10. This sampling process was performed repeatedly, and an average of the resulting correlations was taken. The statistical significance of the result was evaluated using a permutation test as described earlier.

Results

Gene Sequences

Our analyses involved ortholog sets from either four mammalian species or four *Drosophila* species, related as shown in figure 1. We found that these analyses were compromised by the quality of the originally reported genome sequences of the rhesus macaque (Rhesus Macaque Genome Sequencing and Analysis Consortium et al. 2007) and *D. simulans* (Drosophila 12 Genomes Consortium et al. 2007), and therefore used alternative sets of coding sequences for these organisms. For *D. simulans*, we used an assembly of high-coverage RNA-seq reads. For the rhesus macaque, we used a more recent genome sequence (Yan et al. 2011). We found that we could improve upon the annotations of this genome; apparently the propagation of gene models from the earlier, less complete rhesus genome sequence had led to some incorrect gene models (see supplementary information, Supplementary Material online, for an example). We therefore utilized coding sequences generated by our own gene-finding techniques.

ER Correlations between Clades

For each ortholog quartet, we estimated d*N*, d*S*, and d*N*/d*S* for each branch of the phylogenetic tree. As illustrated in figure 2, the correlation between primate and rodent protein ERs (d*N*) is strong, but far from perfect (Spearman's rank-order correlation 0.70). The same holds for the two *Drosophila* species pairs (correlation 0.79). The relationships are similar for d*N*/d*S* (supplementary fig. S5, Supplementary Material online).

One source of deviation from a perfect correlation is sampling variance due to the finite number of substitutions observed. We assessed this effect by simulating Poisson sampling of estimated rates. The inferred substitutions were partitioned between the two lineages according to branch lengths, corresponding to a nonfluctuating ER and a perfect correlation. We then drew Poisson-distributed samples using these substitution numbers as the means. This produced correlations significantly higher than those observed for the real data: 0.87 for mammals and 0.91 for flies. Thus, sampling variance does not explain all of the deviation from a perfect correlation, and ERs indeed vary between lineages. The results presented later provide information about the nature and causes of this rate variation.

Do Factors That Correlate with Rate Also Correlate with Rate Variation?

Based on cross-species comparisons, genes can be classified with respect to "age" or phylogenetic depth, ranging from presence in the last common ancestor of all life to narrow

Table 1
Correlation between ER Fluctuation and Phylogenetic Depth

	Mammals	Flies
Genes	4,330	6,309
dN/dS	−0.05***	−0.01
dN	−0.05***	0.02
dS	−0.04**	−0.08***

**0.001 ≤ P < 0.01.
***P < 0.001.

Table 2
Correlation between ER Fluctuation and Expression Level

	Mammals		Flies	
	EST Count (Human)	EST Count (Human + Mouse)	*D. mel.* mRNA Abundance	*D. mel.* Protein Abundance
Genes	2,976	2,607	8,137	2,981
dN/dS	0.03	0.03	0.06***	0.12***
dN	0.04*	0.06**	0.06***	0.09***
dS	0.00	−0.01	−0.02*	0.07***

*0.01 ≤ P < 0.05.
**0.001 ≤ P < 0.01.
***P < 0.001.

lineage-specificity. Table 1 shows the relationship between phylogenetic depth and variation in ER. For the two measures of relative protein ER, d*N* and d*N*/d*S*, little or no correlation is apparent. Thus, the "age" of a gene tells us little about how much and how rapidly its ER varies.

Table 2 shows the relationship between mRNA or protein abundance and variation in ER. Little or no correlation is observed. What little correlation is observed for flies can apparently be explained largely by greater sampling error for highly expressed proteins because they tend to evolve more slowly, as demonstrated by the sampling analysis shown in table 3. Thus, despite the fact that expression level is the best known predictor of ER, it does not correlate strongly with the tendency of that rate to fluctuate.

Relationship between ER and Its Temporal Variability

If changes to a protein's ER are caused by changes to the protein's sequence, there should be a relationship between the rate at which a protein evolves and the degree to which its ER varies. If, on the other hand, changes in ER are due to factors external to the protein, no such relationship is expected.

We investigated the relationship between the variability of each rate parameter (d*N*, d*N*/d*S*, and d*S*) and the corresponding rate parameter for the internal branch of the phylogenetic tree for each ortholog quartet. The internal branch value served as an estimate of the mean of the rate parameter. This is, we believe, preferable to using the external branch values for this purpose: because these values are used to calculate the ratio, using them for the rate estimate as well could lead to artificial correlations. As illustrated in figure 3, variability was assessed by the absolute deviation of the ratio (primate/rodent or fly1/fly2) from a regression line relating the ratio to the internal branch value in the logarithmic domain. The relationship between rate and variability was measured by the rank-order correlation between the deviations and internal branch parameters.

As shown in table 4, the correlation between protein ER and its variability is weak. For both mammals and *Drosophila*, the correlation coefficients for d*N* and d*N*/d*S*, though statistically distinguishable from zero, are small in magnitude, ranging from −0.09 to −0.17.

Even this weak correlation might be artifactual. Genes with higher d*N* and d*N*/d*S* will tend to have a larger number of observable substitutions and hence lower sampling variance for the rate parameters. This statistical fact would lead to a negative correlation even in the absence of a biological effect. To investigate the contribution of this effect, we performed computations that eliminated it by randomly thinning the substitution counts in the branches used to calculate the ratios. This leaves all genes with the same total number of substitutions in these branches, but with a variable distribution between branches that reflects the ratio of rates. Hence, it largely equalizes sampling variance among proteins with different ERs. For simplicity, this analysis uses protein p-distances based on a most parsimonious reconstruction.

The sampling procedure eliminates much or all of the small negative correlations described earlier (table 5). Thus, those correlations are, in part or in whole, artifacts of sampling variance. The true correlation between protein ER and its tendency to fluctuate appears to be close to zero.

The use of a regression line to remove trends was motivated by the appearance of a rate/ratio correlation for d*S* in flies when rate parameters were simply normalized to the geometric mean. The use of deviations from the overall mean rather than regression residuals has little effect on the correlations for d*N* or d*N*/d*S*. The small negative correlation for sampling with $n = 5$ for flies (table 5) becomes very close to zero, and statistically insignificant, when deviations from the mean are used; apparently the line-fitting and rank-order correlation can interact when *n* is small to produce artifactual correlations.

We also performed a test based on the index of dispersion. A measure of the departure from a Poisson process is the quantity $(R - 1)/M$, where *R* is the index of dispersion and *M* is the mean. This quantity will be independent of the mean ER if the variability of the rate, as measured by the coefficient of variation (ratio of standard deviation to mean), does not depend on the rate. We estimated this quantity from the terminal substitution counts using the method of Gillespie (1989) and calculated its correlation with the internal branch count.

Table 3
Sampling Analysis for Correlation between ER Fluctuation and Expression Level

Sample Size	Mammals						Flies					
	EST Count (Human)			EST Count (Human + Mouse)			*D. mel.* mRNA Abundance			*D. mel.* Protein Abundance		
	Genes	Correlation Coefficient		Genes	Correlation Coefficient		Genes	Correlation Coefficient		Genes	Correlation Coefficient	
		All Substitutions	Sampled Substitutions		All Substitutions	Sampled Substitutions		All Substitutions	Sampled Substitutions		All Substitutions	Sampled Substitutions
5	2,365	0.03	0.00	2,062	0.05*	0.00	6,796	0.07***	0.00	2,532	0.09***	0.00
8	2,421	0.02	0.00	2,103	0.04*	0.01	7,145	0.06***	0.01*	2,653	0.08***	0.02*
15	2,036	0.01	−0.02	1,755	0.03	−0.01	6,383	0.05***	0.02*	2,345	0.06**	0.02
25	1,424	0.01	−0.02	1,211	0.03	0.00	5,018	0.04**	0.03**	1,844	0.01	−0.01
50	663	0.06	0.01	548	0.03	−0.02	2,652	0.04*	0.02	967	0.03	0.01

$*0.01 \le P < 0.05$.
$**0.001 \le P < 0.01$.
$***P < 0.001$.

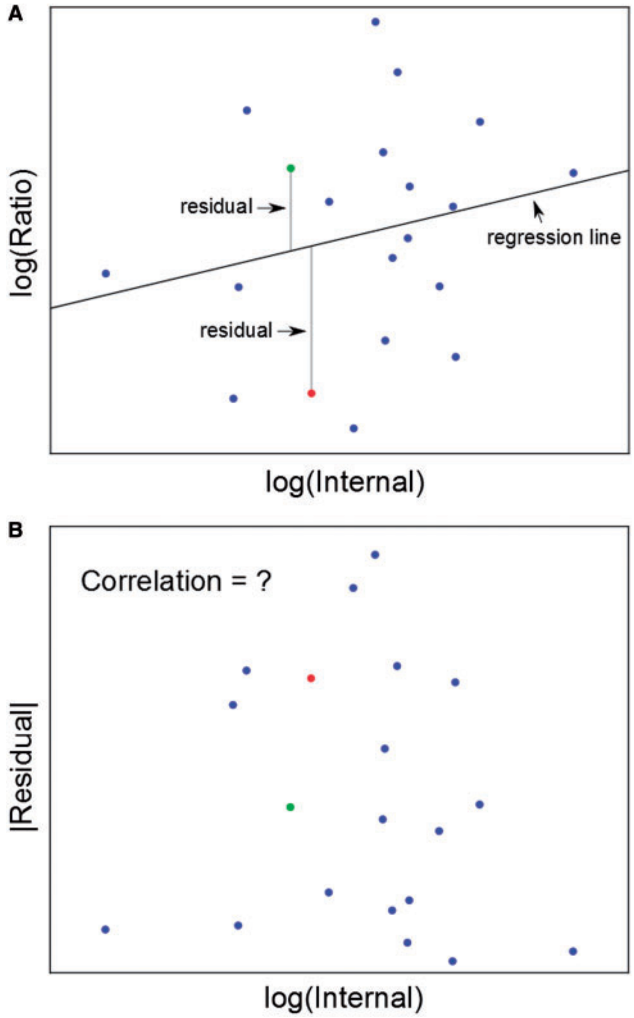

Fig. 3.—Analysis of the relationship between ER and its temporal variation. The method is illustrated using a set of hypothetical genes, each represented by a point. The horizontal axis represents the logarithm of the estimated rate parameter (e.g., d*N*) for the internal branch of the tree. (*A*) The vertical axis represents the logarithm of the ratio of the rate parameter estimates for the two terminal lineages (e.g., primates and rodents). A least-squares line is fit, and the absolute deviations from the line are taken as estimates of rate variability. (*B*) The vertical axis represents the magnitude of these deviations. The coefficient of correlation between the variables in (*B*) is an indication of the relationship between rate and rate variability.

Table 4
Correlation between ER Fluctuation and ER

	Mammals	Flies
Genes	5,053	8,726
d*N*/d*S*	−0.16***	−0.09***
d*N*	−0.17***	−0.10***
d*S*	0.04	0.03*

$*0.01 \le P < 0.05$.
$***P < 0.001$.

Table 5

Sampling Analysis for Correlation between ER Fluctuation and ER

Sample Size	Mammals			Flies		
		Correlation Coefficient			Correlation Coefficient	
	Genes	All Substitutions	Sampled Substitutions	Genes	All Substitutions	Sampled Substitutions
5	4,041	−0.16***	0.00	7,238	−0.10***	−0.05**
8	4,142	−0.14***	−0.03***	7,618	−0.09***	0.00
15	3,479	−0.10***	−0.04***	6,808	−0.07***	0.01
25	2,473	−0.08***	−0.05**	5,365	−0.04**	0.00
50	1,103	−0.04	−0.03	2,844	−0.05*	−0.02

*$0.01 \le P < 0.05$.
**$0.001 \le P < 0.01$.
***$P < 0.001$.

Table 6

Correlations of Rate Changes for Different Domains of a Protein, with Correlations for Adjacent Genes Shown for Comparison

	Mammals			Flies		
	Adjacent Domains	Adjacent Genes	Significance of Difference[a]	Adjacent Domains	Adjacent Genes	Significance of Difference[a]
Cases	393	1,631		496	6,453	
d*N*/d*S*	0.18***	0.04	**	0.14**	0.06***	*
d*N*	0.17***	0.05*	*	0.21***	0.09***	**
d*S*	0.33***	0.34***		0.14**	0.19***	

[a] The statistical significance of the difference between the correlation coefficient for adjacent domains and that for adjacent genes.
*$0.01 \le P < 0.05$.
**$0.001 \le P < 0.01$.
***$P < 0.001$.

For mammals, the Spearman rank-order correlation was 0.101. The Pearson product-moment correlation, however, was indistinguishable from zero (correlation coefficient = −0.008, not statistically significant). For flies, the Spearman correlation was 0.169 but the Pearson correlation was only 0.062. Simulations show that when rate variability does not depend on rate, the Spearman correlation is nonetheless moderately positive, whereas the Pearson correlation is close to zero (supplementary information, Supplementary Material online). We therefore conclude that, at least for mammals, this test also detects no relationship between rate and rate variability.

Covariation of ERs between Different Domains of the Same Protein

Many proteins consist of more than one domain. Domains may fold independently of one another, and they often correspond to functional units. We therefore asked the following question: If the ER of one domain changes over evolutionary time, does that of another domain in the same protein tend to change in the same way, or do the ERs of the different domains vary independently?

To address this question, we estimated ERs for protein domains in pairs of taxa, and compared the ratio of rates for one domain with the ratio for an adjacent domain of the same protein. As shown in table 6, the ERs of adjacent protein domains do tend to fluctuate together: Moderate positive correlations are observed for both flies and mammals, using either d*N* or d*N*/d*S* as the measure of protein ER. Positive correlations are also observed for d*S*.

A possible contributor to these correlations, at least for d*N* and d*S*, is local fluctuation of the mutation rate. More generally, the correlations might reflect the fact that the domain pairs are encoded by nearby parts of the genome rather than the fact that they are part of the same protein. To assess the contribution of this factor, we calculated analogous correlations for pairs of genes that are adjacent on the chromosome. As table 6 shows, the correlations for d*N* and d*N*/d*S* for adjacent genes, though still positive, are significantly smaller than the correlations for domains of the same protein. For d*S*, in contrast, the correlations for adjacent genes are indistinguishable from those for domains in the same protein. This is consistent with some contribution of location-specific effects to the correlations for d*N* and d*N*/d*S* but a comparable or larger contribution from fluctuations in the selective forces acting on protein sequences. These results likely underestimate the contribution of the within-gene effect because individual domains tend to be smaller than whole proteins, leading to higher sampling variance, and hence weaker apparent correlations, for the within-gene ratios.

Although these results establish a between-domain relationship between rate fluctuations, they leave open the possibility that the within-domain relationship is stronger still. We therefore sought to compare within-domain correlations with between-domain correlations. Because distance in the protein sequence might affect the coupling of rates between amino acid positions, we considered within- and between-domain pairs of sequence regions that were separated by the same distance and had the same length (fig. 4). We also report the correlations for the more distant between-domain pairs, which might be lower than within-domain correlations simply because of the greater distance. Table 7 shows that the correlation for regions in different domains is just as strong as the correlation for regions in the same domain: In every case, the correlation coefficients for the three possible pairings of regions are statistically indistinguishable. Even if we relax the requirement for equal spacing of the three regions (supplementary fig. S2, Supplementary Material online), which increases statistical power but might disfavor between-domain correlations, the within-domain correlations are not significantly larger than the between-domain correlations (supplementary table S1, Supplementary Material online). Thus, it appears that rates at positions in different domains vary together just as much as positions in the same domain.

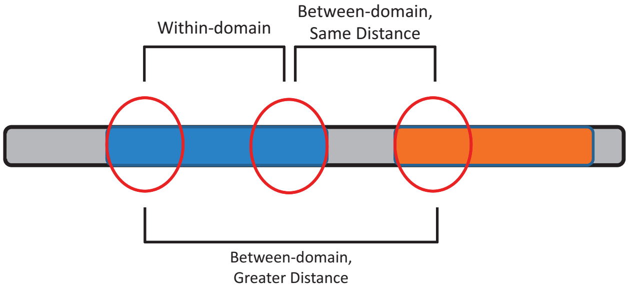

Fig. 4.—Scheme for comparing within- and between-domain covariation of ERs. Two regions in one domain and one in an adjacent domain were chosen so as to maximize their common size, subject to the constraint that they are equally spaced. In some cases, the two regions come from the second domain rather than the first, and the relationships between the region boundaries and the domain boundaries vary with relative domain size and spacing.

Discussion

Fluctuations in the ERs of proteins can provide information about the forces that shape protein evolution. Studies based on the index of dispersion have played a role in the ongoing effort to elucidate the nature of protein sequence evolution (Takahata 1987; Gillespie 1989, 1991; Bedford and Hartl 2008; Bedford et al. 2008). We have used a more straightforward measure of rate variability, based on the ratio of the estimated ERs in different parts of the phylogenetic tree, to test whether various factors correlate with rate variability. In addition, we have investigated the extent to which different domains within a protein vary together in ER, as opposed to acting as isolated units whose rates fluctuate independently.

We found that the variability of a protein's ER does not correlate strongly with phylogenic depth or expression level. This may not be surprising, but is to be contrasted with ER itself, which does correlate with these factors. The lack of correlation is particularly striking for expression level, which universally displays a relatively strong negative correlation with ER.

We found little correlation between a protein's rate of evolution and the amount of temporal variability of that rate. That is, more rapidly evolving proteins did not appear to have more (or less) variable rates of evolution. This result suggests that it is not changes to the protein's own sequence that are primarily responsible for changes to its rate of evolution. It points to the importance of extrinsic factors, such as sequence evolution at other loci and changes to the environment, which can alter the selective forces acting on a protein sequence.

Table 7

Correlations of Rate Change for Pairs of Regions in the Same Domain or in Adjacent Domains

	Mammals (121 Genes)			Flies (147 Genes)		
	Within-Domain	Between-Domain		Within-Domain	Between-Domain	
		Same Distance	Greater Distance		Same Distance	Greater Distance
d*N*/d*S*	0.31***	0.34***	0.34***	0.12	0.07	0.07
d*N*	0.25**	0.20*	0.28**	0.33***	0.18*	0.25**
d*S*	0.26**	0.29***	0.30**	0.03	0.14	0.01

Note.—Within each row the three values for each taxon are statistically indistinguishable.
*$0.01 \leq P < 0.05$.
**$0.001 \leq P < 0.01$.
***$P < 0.001$.

Some caution is warranted when interpreting results of this type. Even if we suppose that all changes to a protein's ER are caused by changes to its sequence, it is not clear how strong a correlation we should expect between measures of rate and measures of rate variability. The value of this correlation depends on the details of the evolutionary process and on the lengths of the branches used in the analysis. Higher ERs lead to more sequence differences between taxa (and hence more variability), but also to greater averaging of rates within taxa (and hence less variability). Thus, even the direction of the correlation that would result is not clear, and any correlations might be small in magnitude and difficult to detect. Conversely, even if rates vary due to factors other than changes to the protein's sequence, a correlation is possible. Simply increasing all selection coefficients by some factor, for example, might affect the ERs of slowly evolving and rapidly evolving proteins to different extents. Furthermore, such factors as sequencing errors, inaccurate gene models, imperfect identification of orthology, and misalignment of sequences have the potential to degrade or bias results.

Bedford et al. (2008) reported a small but nonnegligible relationship between ER and a measure of rate variation in mammals, flies, and yeast. We present a modified analysis of the same data in supplementary information, Supplementary Material online. For mammals, the reported relationship appears to be largely an artifact; our modified analysis yields a correlation close to zero. For *Drosophila*, we infer approximately the same relationship as Bedford et al., though we point out factors that might contribute to an artificial correlation.

We noted earlier that the quality of some genome sequences, and inaccuracies in their annotations, were initially obstacles to this study (see supplementary information, Supplementary Material online, for an example). We believe that we largely overcame these and other sources of error in several ways: constructing our own assemblies or gene models in some cases, filtering alignments based on knowledge of intron/exon structure of the coding sequences, and using different parts of the phylogenetic tree for estimates of mean rate and rate variation (the internal branch for the rate, the terminal branches for rate variation). We would note, however, for both producers and users of sequence data, that the presence of even a small number of inaccurate coding sequences can have a large effect on studies of this type.

The individual domains of a protein correspond to structural units and, to some extent, functional modules. Thus, it might be expected that fluctuations in ER affect domains as somewhat independent units. We found, however, evidence to the contrary. Protein domain boundaries do not appear to be important to the mechanisms that change a protein's ER. The ERs of regions in different domains tend to vary together. This tendency is just as strong for such regions as it is for equally distant regions that are part of the same protein domain. This result is difficult to reconcile with the hypothesis that fluctuation in ER is caused mainly by substitutions that stabilize or destabilize the folded state, since the within-domain effect of such substitutions should be substantially larger than their effect on other domains.

Supplementary Material

Supplementary information, tables S1–S3, and figures S1–S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Richa Agarwala for assembling the *D. simulans* RNA-Seq data and Alexandre Souvorov for producing gene predictions for the Chinese rhesus macaque. They also thank Trevor Bedford and Yuri Wolf for sharing data. This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

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.

Bedford T, Hartl DL. 2008. Overdispersion of the molecular clock: temporal variation of gene-specific substitution rates in *Drosophila*. Mol Biol Evol. 25:1631–1638.

Bedford T, Wapinski I, Hartl DL. 2008. Overdispersion of the molecular clock varies between yeast, *Drosophila* and mammals. Genetics 179: 977–984.

Bloom JD, Raval A, Wilke CO. 2007. Thermodynamics of neutral protein evolution. Genetics 175:255–256.

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

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

Drosophila 12 Genomes Consortium, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218.

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.

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

Gillespie JH. 1989. Lineage effects and the index of dispersion of molecular evolution. Mol Biol Evol. 6:636–647.

Gillespie JH. 1991. The causes of molecular evolution. New York: Oxford University Press.

Kapustin Y, Souvorov A, Tatusova T, Lipman D. 2008. Splign: algorithms for computing spliced alignments with identification of paralogs. Biol Direct. 3:20.

Kreitman M. 1996. The neutral theory is dead. Long live the neutral theory. Bioessays 18:678–683.

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.

Langley CH, Fitch CH. 1974. An estimation of the constancy of the rate of molecular evolution. J Mol Evol. 3:161–177.

Marchler-Bauer A, et al. 2011. CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acids Res. 39:D225–D229.

Ohta T. 1996. The neutral theory is dead. The current significance and standing of neutral and nearly neutral theories. Bioessays 18:673–677.

Ohta T, Kimura M. 1971. On the constancy of the evolutionary rates of cistrons. J Mol Evol. 1:18–25.

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

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35:D61–D65.

Rhesus Macaque Genome Sequencing and Analysis Consortium, et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222–234.

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

Sayers EW, et al. 2011. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 39:D38–D51.

Sayers EW, et al. 2012. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 40:D13–D25.

Schrimpf SP, et al. 2009. Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. PLoS Biol. 7:e48.

Takahata N. 1987. On the overdispersed molecular clock. Genetics 116: 169–179.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science 278:631–637.

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 U S A. 106:7273–7280.

Yan G, et al. 2011. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. Nat Biotechnol. 29:1019–1023.

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

Associate editor: Tal Dagan