Universal distribution of mutational effects on protein stability, uncoupling of protein robustness from sequence evolution and distinct evolutionary modes of prokaryotic and eukaryotic proteins

Guilhem Faure and Eugene V Koonin
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
E-mail: koonin@ncbi.nlm.nih.gov

Abstract
Robustness to destabilizing effects of mutations is thought of as a key factor of protein evolution. The connections between two measures of robustness, the relative core size and the computationally estimated effect of mutations on protein stability (ΔΔG), protein abundance and the selection pressure on protein-coding genes (dN/dS) were analyzed for the organisms with a large number of available protein structures including four eukaryotes, two bacteria and one archaeon. The distribution of the effects of mutations in the core on protein stability is universal and indistinguishable in eukaryotes and bacteria, centered at slightly destabilizing amino acid replacements, and with a heavy tail of more strongly destabilizing replacements. The distribution of mutational effects in the hyperthermophilic archaeon Thermococcus gammatolerans is significantly shifted toward strongly destabilizing replacements which is indicative of stronger constraints that are imposed on proteins in hyperthermophiles. The median effect of mutations is strongly, positively correlated with the relative core size, in evidence of the congruence between the two measures of protein robustness. However, both measures show only limited correlations to the expression level and selection pressure on protein-coding genes. Thus, the degree of robustness reflected in the universal distribution of mutational effects appears to be a fundamental, ancient feature of globular protein folds whereas the observed variations are largely neutral and uncoupled from short term protein evolution. A weak anticorrelation between protein core size and selection pressure is observed only for surface residues in prokaryotes but a stronger anticorrelation is observed for all residues in eukaryotic proteins. This substantial difference between proteins of prokaryotes and eukaryotes is likely to stem from the demonstrable higher compactness of prokaryotic proteins.

Introduction
Protein-coding genes in any organismal lineage evolve at widely different rates, spanning a range of about three orders of magnitude [1, 2]. According to the molecular clock model, each gene is endowed with a characteristc evolutionary rate (ER) that remains approximately constant over long time intervals, up to the entire span of the history of life, in the case of universal genes [3]. However, subsequent measurements have shown that molecular clock is strongly over-dispersed, i.e. the ERs vary to a much greater extent than expected from sampling error alone, under the assumption of a Poisson mutational process [4–7]. The distributions of the rates across sets of orthologous genes in diverse life forms, from bacteria to mammals, which reflect the relative rates of gene evolution, show a notably greater degree of conservation than the absolute rates [1, 2]. This observation inspired the universal pacemaker, a more general model of evolution, that postulates genome-wide, synchronous changes in the ERs of genes. The universal pacemaker model yields a better fit between thousands of individual gene trees and the species tree than the strict molecular clock model [8, 9].
The conservation of the ER distribution implies simple, universal underlying factors. Traditionally, it has been assumed that ER is a multiplicative function of two terms, one of which reflects the intrinsic structural–functional constraints that affect the given protein, whereas the second one corresponds to the biological importance of the same protein [10]. Although this concept is fundamentally plausible and straightforward, for many years, it remained effectively inaccessible to empirical assessment. However, this situation has changed with the advances of functional genomics and systems biology, when prominent correlations have been shown to exist between several evolutionary and molecular phenomic variables [11–15]. 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, non-essential genes evolve slightly faster than essential genes [15–21]. Among all the detected connections, by far the most consistent and strongest one is the universal anticorrelation between the expression level of a gene and its sequence evolution rate: in all model organisms for which detailed expression data are available, highly expressed genes evolve significantly slower than lowly expressed ones [20,22–24].

The universal link between gene expression and sequence evolution inspired the hypothesis that protein abundance or, more precisely, translation rate of protein-coding genes is the key determinant of the sequence evolution rate [23,25,26]. Specifically, the mistranslation-induced misfolding (MIM) hypothesis posits that the major 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 [25–28]. 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 [26]. A direct indication that translation rate substantially contributes to the rate of sequence evolution has been obtained through comparative analysis of the ERs of domains in multidomain proteins that revealed significant homogenization of the ERs compared to the same domain combinations in separate proteins [29]. Quantitative analysis of these results has suggested that the contribution of translation rate was comparable in magnitude to the contribution of structural–functional constraints. Subsequently, a more precise mathematical model has been developed to estimate the relative contributions of structural–functional constraints and translation rate to the rate of evolution of protein-coding genes through a joint comparative analysis of proteomic data on protein abundance and short term ERs in different lineages of animals [30]. The results of this analysis indicate that the contribution of structural–functional constraints is typically 3 to 5 times greater than the contribution of translation rate. Subsequent studies have refined the MIM hypothesis through the demonstration that selection for avoidance of spurious protein–protein interactions additionally contributes to the anticorrelation between expression and ER by constraining the evolution of surface residues in highly expressed proteins [31].

A parallel series of analyses has revealed a strong dependence between the protein core size and the ER of solvent-exposed amino acid residues, and hence complete proteins. It has been shown that increasing the size of the buried protein core results in increased rate of evolution of solvent-exposed residues, which typically constitute the majority of a protein molecule, whereas buried residues are less affected if at all [32–34]. Furthermore, it has been demonstrated that in yeast, expression level and core size exert largely independent effects on the rate of evolution. At the whole protein level, expression has been found to make a substantially greater contribution to the ER than core size although the contributions appeared comparable at the level of individual exposed residues [34,35].

Another line of investigation has demonstrated a positive correlation between protein contact density and the ER such that robust (designable) proteins with a high contact density evolve measurably faster than proteins with a low contact density [36].

Protein stability that is obviously linked to mutational robustness is considered to be one of the key factors of evolution that seems to impose limits on the ER of protein-coding genes [37–40]. The free energy of protein folding is typically in the range of 5–15 kcal mol$^{-1}$ [41]. Thus, proteins appear to be marginally stable so that amino acid replacements that result in free energy increase by over 3 kcal mol$^{-1}$ can be highly destabilizing and hence deleterious [42]. Experimental determination of protein stability remains a relatively difficult task so that genome/proteome scale analysis of the mutational effects on it is currently impractical. However, powerful and apparently accurate computational methods for estimating the free energy of protein globules and the effects of mutations on it have been developed [43–46]. Using one of these methods, FoldX [43,44], Tokuriki et al have shown that the distributions of mutational effects on protein stability were remarkably similar for 21 diverse proteins and closely matched the available experimental data [47].

We sought to investigate the relationships between protein robustness to mutation, abundance and selection pressure on protein-coding genes (a correlate of ER) by taking the maximum advantage of the growing collection of protein structures from diverse model organisms. We come to the conclusion that variation of mutational robustness is largely neutral with respect to the selective pressure on the sequences of protein-
coding genes, especially in prokaryotes. In contrast, in eukaryotes, a moderate dependence exists between the relative protein core size, which is a proxy of robustness, and selection pressure, particularly when measured separately for the surface residues. In addition, we show that proteins from a hyperthermophilic archaeon are on average significantly less robust to mutation than proteins of mesophiles, conceivably due to stronger constraints on protein structure exerted by extreme conditions.

Results

Coverage of genomes of model organisms with protein structures and the universal dependence of \( dN/dS \) on protein abundance

We selected four eukaryotic species (\( S. \) cerevisiae, \( C. \) elegans, \( D. \) melanogaster and \( H. \) sapiens), two bacteria (\( E. \) coli and \( B. \) subtilis) and one archaeon (\( T. \) gammatolerans) for each of which a sufficient number of protein structures have been solved to enable statistical analysis. The coverage of the proteomes with protein structures varied from 0.5% for \( D. \) melanogaster to 19% for \( E. \) coli (table 1; see methods for details). The universal, statistically significant negative correlation between the selection pressure on protein-coding sequences (the inverse of the ratio of non-synonymous to synonymous substitution rates, \( dN/dS \) [48]) versus the protein abundance (expression level) of each protein was detected in all species (table 2; see methods for details), supporting and extending previous findings [20, 22, 26]. It is notable, however, that this dependence was less pronounced in prokaryotes than in eukaryotes. The same correlation was observed also for the sets of proteins with resolved structures although due to the low structure coverage, the correlation coefficient varied within a broader range, and the correlation was not significant anymore for \( T. \) gammatolerans and \( C. \) elegans (table 2).

Effects of mutations in the core on protein stability

For each protein, we explored the distribution of mutation effects in the core positions on the protein stability. Each core position was computationally mutated independently to each of the 19 other amino acids and the \( \Delta \Delta G \) values (i.e. the difference between the \( \Delta G \) values before and after the mutation) were estimated using the FoldX method [44] (see methods for details). The \( \Delta \Delta G \) values were grouped into 1 kcal mol\(^{-1}\) bins from \(-10\) kcal mol\(^{-1}\) to 16 kcal mol\(^{-1}\), and the fraction of mutation resulting in the respective magnitude of \( \Delta \Delta G \) was computed for each bin. We generated and analyzed the distributions of the fractions of mutations associated with different \( \Delta \Delta G \) values for all possible amino acid replacements (19 mutations per position) and separately for replacements resulting from single nucleotide substitutions (9 possible mutations). In the latter distribution, two additional fractions were included, namely (i) silent mutations (mutations leading to the same amino acid) and (ii) nonsense mutations (mutations leading to a stop codon).

To our knowledge, this is the first genome-wide comparative analysis of the distributions of mutation effects on protein stability. In a previous, small scale study, similar distributions have been generated using FoldX and compared with experimental data for 21 proteins. Tokuriki et al have shown that the typical asymmetric distribution of the \( \Delta \Delta G \) values was independent of protein sequence and fold, and that most of the mutations yielded relatively small, positive \( \Delta \Delta G \) values (i.e. slightly destabilized the respective proteins) whereas stabilizing mutations (\( \Delta \Delta G < 0 \)) were rare [47]. The distributions for single nucleotide substitutions included fewer destabilizing mutations and more neutral mutations (mutations with a \( \Delta \Delta G \) around 0 kcal mol\(^{-1}\)) than the complete distributions due to the well-characterized mutational robustness of the genetic code [49, 50].

The universal distribution of mutation effects across proteomes

The mutation effects show highly similar, skewed bell-shaped distributions, with a heavy tail of destabilizing mutations for all analyzed proteomes (figure 1). For the purpose of this comparison, we represent the mutational effects for all proteins in a given species with a boxplot, where the box represents the distribution of fractions of mutations resulting in specific \( \Delta \Delta G \) values for the given proteome. Figure 1 directly compares the distributions of mutation effects for five analyzed proteomes (i.e. \( S. \) cerevisiae, \( H. \) sapiens, \( E. \) coli, \( B. \) subtilis and \( T. \) gammatolerans; \( C. \) elegans and \( D. \) melanogaster proteomes were excluded from this analysis due to the small number of available structures). There was no significant difference between the eukaryotic and bacterial proteomes suggesting that the

---

Table 1. Protein sequence and structure data sets for the analyzed species.

| Species          | S. cerevisiae | D. melanogaster | C. elegans | H. sapiens | E. coli | B. subtilis | T. gammatolerans |
|------------------|---------------|-----------------|------------|------------|---------|-------------|------------------|
| #proteins        | 4449          | 5746            | 8548       | 13 509     | 2133    | 3062        | 1236             |
| #structures/%proteome covered | 186/4.2% | 48/0.8% | 44/0.5% | 512/3.8% | 407/19.1% | 141/4.6% | 97/7.9% |
| Threshold of sequence identity | 100 | 90 | 90 | 100 | 100 | 100 | 70 |
Table 2. Correlations between selective pressure on protein-coding regions, protein abundance and relative core size.

| Variable 1 | Variable 2 | Spearman correlation coefficient |
|------------|------------|----------------------------------|
| Abundance (all) | Abundance (all) | dN/dS |
| S. cerevisiae | 0.51 | -0.21 |
| D. melanogaster | -0.31 | -0.21 |
| C. elegans | -0.022 | -0.21 |
| H. sapiens | -0.21 | -0.21 |
| E. coli | -0.26 | -0.21 |
| B. subtilis | -0.4 | -0.21 |
| T. gammatolerans | -0.41 | -0.21 |
| Partial correlation: Abundance, relative core size | | |
| Core: | 0.21 | 0.22 |
| Surface: | 0.06 | 0.18 |
| Abundance: | 0.19 | 0.22 |
| Partial correlation: Abundance, relative core size | | |
| Core: | -0.09 | -0.12 |
| Surface: | -0.16 | -0.12 |

Note: dN/dS stands for logarithm of the ratio of non-synonymous to synonymous substitution rates which is the measure of selective pressure on protein-coding sequences. The first row shows the correlation between dN/dS and abundance for all protein-coding genes with orthologs from the respective species; the rest of the correlations are for the proteins with resolved structures. Relative core size is the fraction of amino acid residues in the core, i.e. core size divided by protein size. dN/dS_core is the dN/dS value computed from core residues only. dN/dS_surface is the value computed from surface residues only. Significant correlations are shown in bold (p-value <0.05). Relative core size (proportion of core position). Partial correlations were calculated only in case the original correlation was significant.
distribution of the mutation effects on protein stability is universal across the bacterial and eukaryotic domains of life. In contrast, the only archaeon and the only (hyper)thermophile in the data set, *T. gammatolerans* deviated from this universal pattern, with significantly more destabilizing mutations than found in other organisms (figure 1 and supplementary table S1). This feature of the *T. gammatolerans* proteome is compatible with the previous observations of stronger constraints on protein evolution in hyperthermophiles compared to mesophiles [51, 52].

In the entire set of analyzed proteins, there are very few strongly stabilizing mutations ($\Delta\Delta G < -4$ kcal mol$^{-1}$) (figure 1). The effects of the great majority of the mutations (84%) lied between $-3$ and 6 kcal mol$^{-1}$, and 47% were between 0 and 3 kcal mol$^{-1}$. Given that the free energy of protein folding is distributed roughly between 5 and 15 kcal mol$^{-1}$ [41, 42], these findings indicate that most of the mutations are near neutral (have a negligible effect on protein stability) or slightly destabilizing. As already pointed out, the distribution is skewed toward destabilizing mutations but strongly destabilizing mutations ($\Delta\Delta G > 6$ kcal mol) are rare. In *T. gammatolerans*, the fraction of stabilizing mutations, with effects between $-3$ and 0 kcal mol$^{-1}$, is significantly lower than in other analyzed organisms, and conversely, the fraction of strongly destabilizing mutations is significantly greater (Welch two-sample test). The distributions of the effects of single nucleotide substitutions show similar trends (compare figures 1(a) and (b)).
Although the shape of the distribution of mutation effects is universal, individual proteins in each proteome show broad variation, with some being dramatically more robust to mutations than others (figure 2). For instance, in some proteins, more than 30% percent of all mutations are slightly destabilizing, with $\Delta \Delta G$ between 2 and 3 kcal mol$^{-1}$ whereas in other proteins, only about 6% of the mutations have effects in this range. In the following sections, we investigate factors that potentially could explain this variance.

**Effects of mutations on protein stability and core size**

As shown by extensive theoretical analysis, under the mean field approximation, the fraction of stabilizing amino acid replacements (i.e. replacements with $\Delta \Delta G < 0$) can be used as a proxy of protein robustness to mutation [53]. This measure has the advantage of being independent of the actual $\Delta \Delta G$ values estimated by FoldX that can be inaccurate in some cases (Thiltgen and Goldstein 2012), but only on the fraction of these values that is below a particular threshold. In all analyzed organisms, a strong negative correlation was observed between the fraction of stabilizing mutations ($p_{stab}$) and the median of the $\Delta \Delta G$ distribution that can be considered as a quantitative characteristic of the overall mutational effect on the stability of a protein (figure 3).

We then examined the relationships between median $\Delta \Delta G$ and $p_{stab}$ as proxies of the mutational robustness of a protein and other structural, functional and evolutionary features. Highly significant positive correlations were observed between the relative core size, which was calculated as the proportion of buried residues in a given protein, and $\Delta \Delta G$ median for all 5 analyzed proteomes (figure 4). For $p_{stab}$, the expected negative correlation with the relative core size was observed for four of the five analyzed proteomes; however, these correlations were weak and were significant only for S. cerevisiae and E. coli (supplementary figure S1). Furthermore, for T. gammatolerans, a significant positive correlation was observed (see discussion below). A direct proportionality between the mutational robustness of a protein and the relative core size can be expected because stable proteins with large cores can tolerate more destabilizing mutations than less stable proteins with small cores [32, 54]. In evolutionary terms, stable, large-core proteins are not subject to strong selective pressure to minimize the effect of mutations on stability, in contrast to proteins with small cores that are less robust to
destabilizing mutations but are optimized by evolution such that many mutations have little or no destabilizing effect. The observation that $\Delta \Delta G$ median significantly and positively correlates with the relative core size, whereas $p_{stab}$ shows much weaker negative correlations, implies that on genome scale the former reflects protein robustness better than the latter. Accordingly, all further analysis were performed with both measures of protein robustness but we focus primarily on the results obtained with $\Delta \Delta G$ median.

Mutational effects on stability, protein abundance and purifying selection
A comparison of the $\Delta \Delta G$ median to the protein abundance showed no significant correlation for $S.\ cerevisiae$, $H. \ sapiens$ and $T. \ gammatolerans$. In contrast, a significant negative correlation was observed in bacteria (figure 4) where highly abundant proteins appear to have evolved to allow fewer destabilizing mutations (i.e. are less mutationally robust) than low abundance proteins.

The $\Delta \Delta G$ median for each protein was further compared to the $dN/dS$ value, i.e. the strength of purifying selection, for the respective gene. All proteomes showed some positive correlation that, however, was found to be significant only in $H. \ sapiens$ (figure 4). Thus, slow-evolving proteins tolerate fewer destabilizing mutations, i.e. are somewhat less mutationally robust, than faster-evolving proteins. The correlation increased when only surface residues were included in

---

Figure 4. Correlation of the median mutational effect on protein stability with relative core size, protein abundance and selection on protein sequence for five organisms. The columns from left to right show the plots of median $\Delta \Delta G$ versus relative core size $(p_{core})$, protein abundance $(ab)$, $dN/dS$ for complete protein-coding sequence, $dN/dS$ for core residues, and $dN/dS$ for surface residues. Spearman’s rank correlation coefficient $(cor)$ and $p$-value ($p$-value $<0.001$ are set to 0) are reported for each panel. Species abbreviations: BS, $B. \ subtilis$, EC, $E. \ coli$, HS, $H. \ sapiens$, SC, $S. \ cerevisiae$, TG, $T. \ gammatolerans$.
the comparison, and in this case, the correlation became significant also for *E. coli* (figure 4). Apparently, mutations in the core have a greater effect on the stability and evolution of the surface areas of proteins that on the core itself, in accord with previous observations [34, 35].

No significant correlations with either protein abundance or dN/dS were observed for *pstab* (supplementary figure S1). Taken together, these observations unexpectedly indicate that robustness of proteins to mutation inferred from the ΔΔG distribution obtained using FoldX is linked to abundance in bacteria but not in eukaryotes and is largely decoupled from sequence evolution in all tested organisms. As shown previously, the link between ER and robustness could be stronger at the per-residue level than it is on the whole protein level [35, 55, 56].

**Relative protein core size, evolution and abundance of proteins**

The relative core size (proportion of amino acid residues in the core) is a complementary measure of the mutational robustness of a protein that has been reported to be linked to the evolution of protein-coding genes [51, 52]. For each structure, we identified solvent-inaccessible positions, i.e. the core positions that primarily contribute to the protein fold maintenance [57] (see methods). A moderate but significant positive correlation between dN/dS value and proportion of core positions (relative core size) (methods) was observed in *S. cerevisiae*, *D. melanogaster* and *H. sapiens* but not in *prokaryotes* (table 2 and figure 5). This positive correlation was observed also after correction for protein size using partial correlation analysis although in this case, it remained significant only for *S. cerevisiae*. Using the dN/dS values estimated for the core positions only, we found a significant correlation in eukaryotes (except for *C. elegans* and *D. melanogaster*) but again no correlation in prokaryotes. Finally, we estimated the dN/dS only from the surface (exposed) positions and detected a significant positive correlation with the relative core size in all species except for *T. gammatolerans*; the correlation coefficients for the surface residues were greater than those for core residues or complete proteins. Notably, however, the correlation coefficients were on average higher in eukaryotes (i.e. 0.24 for *H. sapiens*, 0.26 for *C. elegans*, 0.38 for *D. melanogaster* and 0.32 for *S. cerevisiae*) than in bacteria (i.e. 0.24 for *B. subtilis* and 0.12 for *E. coli*). Using more restrictive thresholds for solvent accessibility (5% or 0%) yielded similar but somewhat weaker trends than those detected with the 25% (data not shown). These findings imply that increase in the relative core size results in an increased robustness of the respective protein to mutations, and as a consequence, lowers the pressure of purifying selection in eukaryotic proteins. This effect of the core size is most pronounced in the solvent-exposed, surface parts of proteins, and for prokaryotic proteins, this is the only portion for which the (in this case, weak) connection between the core size and selective pressure is detectable at all.

It seems likely that the difference in the effects of the relative core size on protein evolution in prokaryotes and eukaryotes are due to the greater overall compactness of prokaryotic proteins. Indeed, a comparison of the distributions of core sizes in eukaryotes and prokaryotes showed that the mode of the distributions was shifted toward larger relative core size in prokaryotes whereas the tail of small cores was much heavier in eukaryotes (figure 6(a)). Statistical analysis indicates that the difference between the distributions of the relative protein core sizes in prokaryotes and eukaryotes is highly significant (supplementary table S2) whereas between different eukaryotes and prokaryotes the distributions are statistically indistinguishable (figure 6(b)). Thus, it appears likely that the much stronger connection between the relative core size that is observed in eukaryotes compared to prokaryotes results from the ‘fragility’ of eukaryotic proteins with small cores that slows down their evolution, in agreement with recent observations [58].

Notably, there was a significant negative correlation between the relative core size and protein abundance that was substantially more pronounced in bacteria (but not in the archaeon *T. gammatolerans*) than it was in eukaryotes, and in the former case, survived the correction for the universal correlation between abundance and dN/dS (table 2). Conceivably, protein abundance is associated with an increased number of protein–protein interactions which require a large exposed area to be available in a protein globule, and this effect is particularly pronounced in compact bacterial proteins. For eukaryotic proteins, the selective pressure correlated with the relative core size much stronger than with the median ΔΔG. This difference potentially could be attributed to low accuracy of ΔΔG estimation by FoldX. However, given that such a difference was not observed for prokaryotic proteins, it appears more likely that the relative core size reflects aspects of protein robustness that are distinct from those reflected in the median ΔΔG.

**Structural and evolutionary trends in individual sets of orthologous proteins**

We identified 24 pairs of orthologous proteins from the analyzed organisms with structures available for both orthologs (see methods for details). Despite the small size of this data set, several trends were detectable. In particular, we found that bacterial proteins tend to be more compact, subject to stronger purifying selection and have more stabilizing mutations than their orthologs from eukaryotes. Among the six orthologous pairs that involve a bacterial and a eukaryotic protein, five bacterial structures have a larger relative core size (supplementary table S3).
Thus, in agreement with the results of the bulk analysis described in the preceding section, the prokaryotic proteins are more compact than their eukaryotic counterparts which tend to possess extra decorations that are often involved in transient interactions with other molecules. In the same five cases, the bacterial
proteins are subject to stronger purifying selection (lower \( dN/dS \) values) than their orthologs from eukaryotes (supplementary table S3). A further dissection showed that the difference was concentrated in surface residues that evolve notably slower in bacterial compared to eukaryotic proteins; in contrast, no such trend was detected in the protein cores. We also compared the \( \Delta \Delta G \) medians between for pairs of orthologs and found that in four cases of six, the spectrum of potential mutations in bacterial proteins included more stabilizing and fewer destabilizing mutations (i.e. lower median \( \Delta \Delta G \)) than their respective eukaryotic orthologs (supplementary table S3).

Orthologous proteins typically show closely similar distributions of mutation effects on stability (supplementary figure S2). For each pair of orthologs, we estimated the fraction of the distributions in the respective proteomes that showed the same or greater level of similarity (see methods for details). For 18 ortholog pairs, the similarity level was within the top 20% of the most similar mutability landscapes (supplementary table S3). However, in only two cases (20 S proteasome type 2 protein in yeast and human, and the Copper–Zinc superoxide dismutase in yeast and human), the distributions for the orthologs were closer to one another than to any other distribution between pairs of proteins from the respective proteomes.

Finally, we compared the difference between the distributions of the mutational effects in orthologs to the sequence identity and structural deviation (root mean square deviation (RMSD)). The difference between the distributions in pairs of orthologous protein was not significantly correlated to the sequence identity (supplementary figure S3). These observations on individual pairs of orthologs are compatible with the findings described in the preceding section, with respect to substantial decoupling of variation in mutational effects on protein stability from sequence evolution.

**Discussion**

What determines the rate of protein evolution (or more precisely, selection on protein sequences), is a storied, complex and obviously difficult problem. Clearly, protein abundance (or translation rate) is...
linked to the sequence evolution rate but can hardly be considered the single or even the most important determinant [15, 59]. Efforts on quantitative dissection of the effects of the rate of translation and domain-specific structural and functional constraints suggest that the contributions of both factors are important but the latter is measurably greater [30]. What is the nature of these constraints on protein evolution?

Here, we took advantage of the steadily increasing database of protein structures from diverse organisms to explore some plausible candidates, in particular mutational robustness of protein domains. If protein misfolding is a key component of the fitness cost of mutations, a direct connection should be expected to exist between the robustness of protein folds to mutation and protein stability, on the one hand, and abundance, evolution rate and evolvability of proteins, on the other hand [26, 28, 30, 39, 40]. Despite the striking progress in experimental methods for sequence manipulation, protein expression and especially experimental measurement of protein stability still lag far behind, making genome-wide analysis of protein robustness impractical for any foreseeable future. Accordingly, at present, this problem can only be addressed using computational simulation of protein folding. FoldX employs an empirical force field that was developed to rapidly calculate the free energy (ΔG) of proteins and evaluate the effect of amino acid replacements on ΔG and accordingly the stability of proteins; this effect is measured as ΔΔG, i.e. the difference between the free energies of the native and mutated states [43, 44]. The limited available comparisons with experimental data seem to attest to a reasonable accuracy of the computational estimates [47], and the universality of the ΔΔG distributions observed here is compatible with this assessment. Therefore, in the absence of practicable direct approaches, FoldX appears to be the method of choice for a genome-wide analysis of protein robustness and stability and their evolutionary correlates.

The first result of this analysis to be emphasized is the significant positive correlation between the median ΔΔG of a protein with its relative core size. This correlation seems to indicate that mutational robustness of proteins increases with the enlargement of the core. In other words, proteins with a large proportion of amino acid residues in the core are not ‘fragile’ and thus have not evolved to minimize the effect of mutations. Given that the core size and ΔΔG are determined from the protein structures independently, this finding appears indicative of the structural-functional relevance of the ΔΔG estimates. Larger protein cores can accommodate more destabilizing mutations than smaller cores, i.e. are associated with a greater mutational robustness of proteins. This definition of robustness might seem somewhat counter-intuitive given the greater effect of destabilizing mutations on protein stability but the strong correlation between the median ΔΔG and relative core size implies that it reflects the realities of protein evolution.

The distribution of the ΔΔG values for protein cores was found to be universal among eukaryotes and bacteria, with the mode and median at slightly destabilizing amino acid replacements and a heavy tail of more strongly destabilizing mutations. These findings indicate that the shape of the free energy landscape of the core is a fundamental feature of globular proteins, with similar levels of mutational robustness attained by protein evolution in a broad variety of life forms. This distribution joins the growing list of universal patterns identified by quantitative comparative genomics [60]. By implication, this distribution, with the characteristic level of optimization, can be inferred to have evolved at an early stage of the evolution of each widespread protein fold, probably before the divergence of the three domains of cellular life. This conclusion is compatible with the observations that radiation of the common protein folds into major families antedates the last common ancestor of the extant cellular life forms [61]. However, the universal ΔΔG distribution appears to be significantly shifted towards destabilizing mutations in hyperthermophiles (pending validation by analysis of additional proteomes from hyperthermophilic organisms), resulting in a greater characteristic robustness that seems to reflect the stronger stability-determined constraints on proteins of thermophiles compared to those of mesophiles [51, 52].

Unexpectedly, in most cases, no significant connection was found to exist between the median ΔΔG (a proxy of mutational robustness) and the pressure of purifying selection (dN/dS) affecting a protein sequence (and accordingly, the rate of protein evolution) although the sign of the observed weak correlation was as expected under the hypothesis that more robust (less fragile) proteins are subject to weaker constraints than less robust ones. The connection between mutational robustness and protein abundance is also weak and is significant only in bacteria. Conversely, the relative core size, which is directly linked to robustness, showed significant correlation with dN/dS only in eukaryotes although a stronger dependence was detected to exist for surface amino acid residues in all organisms. For eukaryotic proteins the relative core size was found to be a much better predictor of selective pressure (dN/dS) than it was for prokaryotic proteins. In part, this difference could be attributed to the inaccuracy of FoldX but the lack of a similar difference in prokaryotes implies that the relative core size and median ΔΔG reflect distinct aspects of protein robustness. The nature of this difference remains to be elucidated.

Taken together, these findings seem to indicate, perhaps counter-intuitively, that robustness to destabilizing effects of mutation is not a major, defining factor of protein evolution, at least on the relatively short scale that is suitable for the dN/dS measurements and
on the level of whole proteins. Conceivably, as suggested above, for the common protein folds, the degree of robustness of the cores that is essential for protein stability and function was attained at early stages of evolution whereas much of the subsequent variation appears to be effectively neutral. This conclusion holds particularly well for the compact, highly optimized protein of prokaryotes that generally are subject to strong purifying selection linked to the characteristic large effective population size [62] and show some dependence on core robustness only for surface residues. In the less compact, typically more fragile eukaryotic proteins, the connection between robustness and sequence evolution is significantly stronger.

Conclusions

In this work, we took advantage of the growing collection of protein structures from diverse organisms to examine, on the whole genome scale, the connections between biophysical characteristics of proteins and their evolution. We find that the distribution of the mutational effects on protein stability is a universal characteristic of cellular life forms, with significant deviation detected only in a hyperthermophilic archaeon. The results of this analysis imply that a protein’s robustness to mutation, at least on the whole protein level, is, to a large extent, uncoupled from its abundance and sequence evolution. Such uncoupling appears to be particularly pronounced in the compact, apparently highly optimized proteins of prokaryotes. The less compact, more fragile eukaryotic proteins show a notably stronger connection between the relative protein core size, taken as a measure of mutational robustness, and sequence evolution rate, particularly at the protein surface. The general conclusion from this analysis is that the minimum required or optimal mutational robustness was reached at early stages of protein fold evolution whereas much of subsequent variation was neutral.

Materials and methods

Data sets

Gene orthology

Orthologous sequences of S. cerevisiae were extracted from SensuStricto database which contains S. paradoxus, S. mikatae, S. kudriavzevii and S. bayanus var. uravum [63]. Orthologous pairwise sequences were extracted from OMA database for the following pairs of species: C. elegans and C. briggsae, D. melanogaster and D. pseudoobscura, H. sapiens and M. musculus, B. subtilis and B. subtilis subsp. Spizizenii, E. coli and S. typhi [64]. The orthologous sequences of T. gammatolerans in T. AM4_uid54735 were extracted from the ATGC database [65, 66].

Abundance

Protein abundances were extracted from the Pax database [67] using the available integrated datasets http://pax-db.org/dao/4932-S.cerevisiae_whole_organism-integrated_dataset.txt for S. cerevisiae, http://pax-db.org/dao/6239-C.elegans_whole_organism-integrated_dataset.txt for C. elegans, http://pax-db.org/dao/7227-D.melanogaster_whole_organism-integrated_dataset.txt for D. melanogaster, http://pax-db.org/dao/9606-H.sapiens_whole_organism-integrated_dataset.txt for H. sapiens, http://pax-db.org/dao/511145-E.coli_whole_organism-integrated_dataset.txt for E. coli and http://pax-db.org/dao/224308-Spectral_counting_B.subtili_Chi_MCP_2011.txt for B. subtilis. The logarithm of the protein abundance was taken and scaled between 0 and 1 for each species top obtain the relative abundance.

Methods

Evolutionary analysis

Orthologous proteins were identified from SensuStricto, OMA and ATCG databases (see datasets) and were aligned using the MUSCLE software [68]. Then, codon alignments were constructed by backtracking from the amino acid sequence alignments. The dN and dS values were computed with PAML software [69] using the Maximum Likelihood method by pairs for all species and integrating the phylogenetic tree for S. cerevisiae from SensuStricto. The dN and dS values for the archaeon T. gammatolerans were directly extracted from the current release of the ATCG database [66]. We excluded unreliable estimates resulting from high (mutational saturation) or low number of substitutions. The dN and dS values were corrected as follows: dN or dS superior to 3 are set as 3. dN or dS inferior to 0.001 are set as 0.001. Proteins with both dN and dS above or below the threshold were discarded.

Identification of orthologs between distant species

All protein sequences of S. cerevisiae, H. sapiens, E. coli, B. subtilis and T. gammatolerans were clustered into groups sharing at least 20 percent of sequence identity using usearch software [70]. Only groups of putative orthologs that included proteins covered by structural data and with at least one Eukaryotic and Prokaryotic protein were retained. Each group was manually curated to eliminate potential alignment errors, resulting in a final set of 24 pairs of structurally characterized orthologous proteins.

Protein structures

Proteins with resolved structures were identified by comparing the respective sequences to the Protein Data Bank using BLASTP [71]. Given the insufficient number of structures that are available directly from the model species, structures of proteins that shared
100%, 90% or 70% sequence identity with the query sequence were selected depending on the species. In particular, alternative structures from close species were collected for D. melanogaster, C. elegans (sequence identity 90%) and T. gammatolerans (sequence identity 70%). Only crystal structures with a resolution below or equal to 5 Å were used. We extracted only the protein chain required from the PDB file removing the other protein chains such as protein complexes, solvent, ions, DNA, RNA and small molecules. Side chain atomic coordinates were rebuilt using SCWRL software [72] when at least one side chain atom was missing. Protein structures containing missing backbone atoms were discarded. Structural deviation between orthologous proteins were computed as the RMSD using the ProFit software [73] [74] from the heavy backbone atoms (i.e. Ca, O and N).

Core positions
Solvent accessibility of amino acid residues in protein structures was computed using the NACCESS software [75]. Briefly, buried residue are identified using the relative solvent accessibility (RSA) that is calculated as the percentage of accessibility of a residue in the protein compared to the accessibility of the same residue in the X position of an extended AXA tripeptide [75]. The relative core size (pcore) was defined as the proportion of buried residues in the protein [75].

A position was classified as buried, or belonging to the protein core, when the overall atoms composing the side chain had a RSA below 25%. Although defining the optimal accessibility threshold is a complex problem [76], the 25% threshold has been shown to best partition the residues that constitute the protein core (or protein interior) from surface residues using the specific amino acids compositions in protein cores and surfaces [57]. When several structures overlapped the same position, the position was considered buried only if buried in all structures. In order to restrict the analysis to globular proteins, proteins with fewer than five buried positions were discarded. The proportion of core positions, or the (relative) core size was calculated as the ratio between the number of core position (buried residues) to the total number of amino acid in the given protein.

\[ \Delta \Delta G \]

The \( \Delta \Delta G \) value associated with each mutation was calculated using the FoldX method [44]. Protein structures were relaxed in the FoldX force field after which all positions were mutated one by one and the new free energy value was recorded. The \( \Delta \Delta G \) is the difference between wild type FoldX energy and the mutant FoldX energy. The \( \Delta \Delta G \) values calculated using FoldX were corrected using the PCA equation \( \Delta \Delta G(\text{FoldX}) = -0.078 + 1.14 \Delta \Delta G \) (experimental) [47].

Mutations resulting in negative \( \Delta \Delta G \) values were considered stabilizing and mutations resulting in positive \( \Delta \Delta G \) values were considered destabilizing mutations. When two or more structures covered a same position, the mean \( \Delta \Delta G \) value for each mutant in each position was used.

Distributions of mutational effects on protein stability
The distributions of mutational effects on protein stability mutability landscape were computed following the protocol of Tokuriki et al [47]. The \( \Delta \Delta G \) values were assorted into 1 kcal mol\(^{-1}\) bins from –10 to 16 kcal mol\(^{-1}\) and the fraction of mutation in each was calculated. The comparison of fractions of mutations in each bin across species was performed using an ANOVA and Tukey’s Honest Significance tests. To compare the distributions for orthologous proteins (protein A and protein B from species a and species b, respectively), we first computed the average difference of fraction in mutability landscape (adml) which corresponds to the average difference between pairs of \( \Delta \Delta G \) fractions. The same mean differences were computed between protein A and all distributions for proteins of species B, and the rank of protein A in the resulting distribution of differences was computed.

Acknowledgments
The authors thank Yuri I Wolf for critical reading of the manuscript and numerous helpful suggestions, and the Koonin group members for useful discussions. The authors’ research is supported by intramural funds of the US Department of Health and Human Services (to the National Library of Medicine).

Author contributions
GF performed data analysis, GF and EVK analyzed the results, EVK wrote the manuscript that was read and approved by both authors.

References
[1] Grishin N V, Wolf Y I and Koonin E V 2000 From complete genomes to measures of substitution rate variability within and between proteins Genome Res. 10 991–1000
[2] Wolf Y I, Novichkov P S, Karev G P, Koonin E V and Lipman D J 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 7273–80
[3] Zuckerkandl E and Pauling L 1965 Evolutionary divergence and convergence of proteins Evolving Gene and Proteins ed V Bryson and H J Vogel (New York: Academic) 97–166
[4] Takahata N 1987 On the overdispersed molecular clock Genetics 116 169–79
[5] Cutler D J 2000 Understanding the overdispersed molecular clock Genetics 154 1403–17
[6] Wilke C O 2004 Molecular clock in neutral protein evolution BMC Genetics 5 25
[7] Bedford T and Hartl D L 2008 Overdispersion of the molecular clock: temporal variation of gene-specific substitution rates in drosophila Mol. Biol. Evol. 25 1631–8
[8] Snir S, Wolf Y I and Koonin E V 2012 Universal pacemaker of genome evolution PLoS Comput. Biol. 8 e1002785
[9] Snir S, Wolf Y I and Koonin E V 2014 Universal pacemaker of genome evolution in animals and fungi and variation of evolutionary rates in diverse organisms Genome Biol. Evol. 6 1268–78
[10] Wilson A C, Carlson S S and White T J 1977 Biochemical evolution Annu. Rev. Biochem. 46 573–639
[11] Herbeck J T and Wall D P 2003 Converging on a general model of protein evolution Trends Biotechnol. 23 485–7
[12] Koonin E V and Wolf Y I 2006 Evolutionary systems biology: links between gene evolution and function Curr. Opin. Biotechnol. 17 481–7
[13] Pal C, Papp B and Lercher M J 2006 An integrated view of protein evolution Nat. Rev. Genetics 7 337–48
[14] Vitek D, Kharchenko P and Wagner A 2006 Influence of metabolic network structure and function on enzyme evolution Genome Biol. 7 R39
[15] Wolf Y I 2006 Coping with the quantitative genomics ‘elephant’: the correlation between the gene dispensability and evolution rate Trends Genetics 22 354–7
[16] Duret L and Mouchiroud D 2000 Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate Mol. Biol. Evol. 17 68–74
[17] Hurst L D and Smith N G 1999 Do essential genes evolve slowly? Curr. Biol. 9 747–50
[18] Hirsh A E and Fraser H B 2001 Protein dispensability and rate of evolution Nature 411 1046–9
[19] Jordan I K, Rogozin I B, Wolf Y I and Koonin E V 2002 Essential genes are more evolutionarily conserved than nonessential genes in bacteria Genome Res. 12 1962–8
[20] Krylov D M, Wolf Y I, Rogozin I B and Koonin E V 2003 Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution Genome Res. 13 2229–35
[21] Wall D P, Hirsh A E, Fraser H B, Kumm J, Giaever G, Wolf M Y, Wolf Y I and Koonin E V 2008 Comparable coding-sequence evolution within the context of residue solvent accessibility BMC Evol. Biol. 12 179
[22] Franzosa E A and Xia Y 2012 Independent effects of protein core size and expression on residue-level structure–evolution relationships PLoS ONE e46602
[23] Zhou T, Drummond D A and Wilke C O 2008 Contact density affects protein evolutionary rate from bacteria to animals J. Mol. Evol. 66 395–404
[24] Zeldovich K B, Chen P and Shakhnovich E I 2007 Protein stability imposes limits on organism complexity and speed of molecular evolution Proc. Natl Acad. Sci. USA 104 16152–7
[25] Zeldovich K B and Shakhnovich E I 2008 Understanding protein evolution: from protein physics to Darwinian selection Annu. Rev. Phys. Chem. 59 105–27
[26] Guerisot R, Nielsen J E and Serrano L 2002 Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations J. Mol. Biol. 320 369–87
[27] Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F and Serrano L 2007 The FoldX web server: an online force field Nucleic Acids Res. 35 W382–8
[28] Parthiban V, Gromiha M M, Abhinandan M and Schomburg D 2007 Computational modeling of protein mutation stability: analysis and optimization of statistical potentials and structural features reveal insights into prediction model development BMC Struct. Biol. 7 54
[29] Masso M and Vaisman I 2008 Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis Bioinformatics 24 2002–9
[30] Tokuriki N, Stricher F, Schymkowitz J, Serrano L and Taw C O 2007 The Ka/Ks ratio: diagnosing the form of sequence evolution Trends Genetics 18 486
[31] Knight R D, Freeland S J and Landweber L F 1999 Selection, interactions and chemistry: the three faces of the genetic code Trends Biomed. Sci. 24 241–7
[32] Koonin E V and Novozhilov A S 2009 Origin and evolution of the genetic code: the universal enigma IUBMB Life 61 99–111
[33] Englund J L, Shakhnovich B E and Shakhnovich E I 2003 Natural selection of more desirable folds: a mechanism for thermophilic adaptation Proc. Natl Acad. Sci. USA 100 8727–31
[34] Zeldovich K B, Berezovsky I N and Shakhnovich E I 2007 Protein and DNA sequence determinants of thermophilic adaptation PLoS Comput. Biol. 3 e5
[35] Wilke C O, Bloom J D, Drummond D A and Raval A 2005 Predicting the tolerance of proteins to random amino acid substitution Biophys. J. 89 3714–20
[36] Bloom J D, Silberg J J, Wilke C O, Drummond D A, Adami C and Arnold F H 2005 Thermodynamic prediction of protein neutrality Proc. Natl Acad. Sci. USA 102 606–11
[37] Bloom J D and Glassman M J 2009 Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin PLoS Comput. Biol. 5 e1000349
[38] Echave J, Jackson E L and Wilke C O 2014 Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites (http://bioxiv.org/content/early/2014/09/19/094425)
[57] Levy E D 2010 A simple definition of structural regions in proteins and its use in analyzing interface evolution J. Mol. Biol. 403 660–70
[58] Assis R and Kondrashov A S 2014 Conserved proteins are fragile Mol. Biol. Evol. 31 419–24
[59] Wolf Y I, Carmel L and Koonin E V 2006 Unifying measures of gene function and evolution Proc. Biol. Sci. 273 1907–15
[60] Koonin E V 2011 Are there laws of genome evolution? PLoS Comput. Biol. 7 e1002173
[61] Aravind L, Mazumder R, Vasudevan S and Koonin E V 2002 Trends in protein evolution inferred from sequence and structure analysis Curr. Opin. Struct. Biol. 12 392–9
[62] Lynch M 2007 The frailty of adaptive hypotheses for the origins of organismal complexity Proc. Natl Acad. Sci. USA 104 8597–604
[63] Scannell D R, Zill O A, Rokas A, Payen C, Dunham M J, Eisen M B, Rine J, Johnston M and Hittinger C T 2011 The awesome power of yeast evolutionary genetics: new genome sequences and strain resources for the saccharomyces sensu stricto genus G3 (Bethesda) 111–25
[64] Altenhoff A M, Schneider A, Gonnet G H and Dessimoz C 2011 OMA 2011: orthology inference among 1000 complete genomes Nucleic Acids Res. 39 D289–94
[65] Novichkov P S, Rattrere I, Wolf Y I, Koonin E V and Dubchak I 2009 ATGCC: a database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes Nucleic Acids Res. 37 D448–54
[66] Puigbo P, Lobkovsky A E, Kristensen D M, Wolf Y I and Koonin E V 2014 Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes BMC Biol. 12 66
[67] Wang M, Weiss M, Simonovic M, Haertinger G, Schrimpfs P, Hengartner M O and von Mering C 2012 PaxDb, a database of protein abundance averages across all three domains of life Mol. Cell. Proteomics 11 492–500
[68] Edgar R C 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput Nucleic Acids Res. 32 1792–7
[69] Yang Z 2007 PAML 4: phylogenetic analysis by maximum likelihood Mol. Biol. Evol. 24 1586–91
[70] Edgar R C 2010 Search and clustering orders of magnitude faster than BLAST Bioinformatics 26 2460–1
[71] Altschul S F, Madden T L, Schaffer A A, Zhang J, Zhang Z, Miller W and Lipman D J 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs Nucleic Acids Res. 25 3389–402
[72] Wang Q, Canutescu A A and Dunbrack R I Jr 2008 SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling Nat. Protocols 3 1832–47
[73] McLachlan A D 1982 Rapid comparison of protein structures Acta Crystallogr. A 38 871–3
[74] Martin A C R ProFit v3.1 2009 (http://www.bioinf.org.uk/software/profit/)
[75] Hubbard S J and Thornton J M 1993 NACCESS (University College London)
[76] Tien M Z, Meyer A G, Sdykova D K, Spielman S J and Wilke C O 2013 Maximum allowed solvent accessibilities of residues in proteins PLoS ONE 8 e80635