Determinants of the rate of protein sequence evolution

Jianzhi Zhang and Jian-Rong Yang

Abstract | The rate and mechanism of protein sequence evolution have been central questions in evolutionary biology since the 1960s. Although the rate of protein sequence evolution depends primarily on the level of functional constraint, exactly what determines functional constraint has remained unclear. The increasing availability of genomic data has enabled much needed empirical examinations on the nature of functional constraint. These studies found that the evolutionary rate of a protein is predominantly influenced by its expression level rather than functional importance. A combination of theoretical and empirical analyses has identified multiple mechanisms behind these observations and demonstrated a prominent role in protein evolution of selection against errors in molecular and cellular processes.

Neutral theory
A theory of molecular evolution asserting that most variations of DNA and protein sequences within and between species are selectively neutral rather than adaptive.

Functional constraint
The extent to which random mutations are purged by natural selection owing to their deleterious effects on protein function.

Functional importance
The fitness advantage to an organism conferred by the function of a protein.

Molecular clock hypothesis
The hypothesis that the same protein evolves with an approximately constant rate over time and across different organisms.

Department of Ecology and Evolutionary Biology, University of Michigan, 830 North University Avenue, Ann Arbor, Michigan 48109, USA.
Correspondence to J.Z.
e-mail: jianzhi@umich.edu
doi:10.1038/nrg3950
Published online 9 June 2015

The determination of the amino acid sequences of several homologous proteins in the late 1950s and early 1960s was quickly followed by studies that estimated the rate of protein sequence evolution in different species[1–3]. The rate of protein sequence evolution has remained a central subject in evolutionary and molecular biology for half a century and is critical to the reconstruction of evolutionary history and mechanisms[4,5]. Early studies found that different proteins from the same species can evolve at vastly different rates[2]. According to the well-accepted explanation by the neutral theory[6], the rate of protein sequence evolution (k) equals the rate of mutation (μ) multiplied by the proportion (p) of mutations that are neutral, as beneficial mutations are considered too rare to affect the rate of protein evolution. In theory, p is determined by the functional constraint on the protein: the stronger the functional constraint, the lower the value of p. Although the role of μ in determining k has been clearly demonstrated[7], what constitutes functional constraint has not been clearly defined. As a result, studies have only indirectly estimated the level of functional constraint through the protein evolutionary rate. This circularity hampers mechanistic understanding of protein evolution. In the past 15 years, the increased availability of genomic data for species across the tree of life prompted an extensive search for the major determinants of the protein evolutionary rate. Surprisingly, the functional importance of a protein, widely thought to approximate the level of functional constraint, has only a minor role[8], whereas protein expression level is found to be a major determinant[9]. Subsequent theoretical and empirical studies identified multiple reasons behind the impact of expression level on the rate of protein sequence evolution[10–17]. These discoveries identified an unexpected role in protein evolution of natural selection against errors in molecular and cellular processes.

We review here the main discoveries made in this journey to characterize the rate of protein evolution. We detail several primary hypotheses and models that have been proposed to explain protein evolution rate and mechanism. We synthesize the new mechanistic understanding of protein evolution made possible by recent studies based on analyses of large genomic data sets, and offer our views on the substantial biological and medical implications of the progress made in this area. We focus on the evolutionary rate variation among proteins rather than that among sites within a protein, which is reviewed in REF. 18. We also do not discuss the rate variation of a given protein among different species[7].

Foundations of the field

Early studies examining the rate of protein evolution resulted in two major discoveries that formed the foundations of the fields of molecular evolution and comparative genomics. First, Zuckerkandl and Pauling[2] proposed the molecular clock hypothesis based on findings of an approximately constant rate of evolution for a given protein across different evolutionary lineages. This discovery enabled molecular dating of evolutionary events that could not or did not leave adequate fossil records and now has as important a role as paleontology in providing a temporal scale of biological evolution[19].

Box 1 | Measuring the rate of protein sequence evolution

The rate of protein sequence evolution (*k*) is commonly estimated by the number of amino acid substitutions per site between a pair of orthologous proteins (*d*), divided by twice the time since the divergence between the two species from which the proteins are found (*t*). The simplest method to estimate *d* is to align the orthologous proteins and compute the fraction of aligned amino acid positions that differ between the two sequences. Because not all amino acid substitutions that have occurred in the divergence of the orthologous proteins are observable, elaborated methods of estimating *d* by correcting for unobserved substitutions have been developed and are widely used[90]. The time since divergence (*t*) is commonly inferred from fossil records or estimated from molecular dating. When different proteins from the same species pair are compared, *d* may be directly compared because *t* is the same for all the proteins under consideration. If the interprotein variation in evolutionary rate caused by mutation rate heterogeneity is not of interest and needs to be excluded, one may use *d*N/*d*S as a measure of protein evolutionary rate, where *d*N is the number of nonsynonymous nucleotide substitutions per nonsynonymous site, and *d*S is the number of synonymous nucleotide substitutions per synonymous site[90]. Because *d*S is primarily determined by mutation rate, whereas *d*N is determined jointly by mutation rate and selection, *d*N/*d*S is determined by selection only.

Second, by calculating the evolutionary rates of three proteins, Kimura[3] noticed that the molecular evolutionary rate is too high to have been driven by positive Darwinian selection, which, in conjunction with other observations in molecular biology[20] and population genetic theories, led to the development of the neutral theory[6], the only paradigm-changing conceptual revolution in evolutionary biology since the maturation of neo-Darwinism in the 1950s[21]. The neutral theory asserts that the vast majority of intraspecific polymorphisms and interspecific differences in protein sequence are selectively neutral rather than adaptive, contrasting the view of neo-Darwinists that most intraspecific and interspecific variations are adaptive. Commonly used methods for estimating the rate of protein sequence evolution are explained in BOX 1.

Functional importance

Historical development. The functional importance of a protein refers to the fitness advantage to an organism provided by the function of the protein. It is generally thought that the functional importance of a protein is a major determinant of its evolutionary rate; the more important a protein is, the slower it evolves[22]. This widely accepted belief is probably attributable to an influential article by Kimura and Ohta[23] more than 40 years ago that summarized five principles governing molecular evolution. The second of the five principles reads "functionally less important molecules or parts of a molecule evolve faster than more important ones". At that time, the evolutionary rates of more than 20 proteins had been estimated. The highest evolutionary rate for a protein was observed in fibrinopeptides and was >1,000 times greater than the lowest rate, which was found for histone IV. This comparison supported the notion that more important proteins evolve more slowly because Kimura and Ohta noted that fibrinopeptides have little known function after they become separated from fibrinogen in the blood clot, whereas histones have crucial roles in gene regulation. Nonetheless, they also explained that fibrinopeptides evolve rapidly because nearly every amino acid is acceptable at each position of fibrinopeptides as long as it does not interfere with the cleavage of the peptides, whereas in histones most amino acids at many sites are probably 'unacceptable' because they would affect histone function[23]. Interestingly, this latter explanation was actually referring to a higher functional constraint on histones than fibrinopeptides rather than greater functional importance for the former than the latter. These authors were apparently using the terms functional importance and functional constraint interchangeably despite key differences in these concepts. In comparison to functional importance defined above, functional constraint of a protein refers to the extent to which random mutations are purged by natural selection owing to their deleterious effects on the protein function. In 2009, one of us (J.Z.) asked Ohta in person whether functional importance or functional constraint was in her mind at the time of writing the paper published in 1974. She did not answer immediately but replied 10 days later in an e-mail that it was functional constraint.

Wilson *et al.*[24] made one of the first clear distinctions between functional constraint and functional importance. They suggested that the evolutionary rate of a protein is a mathematical function of dispensability — the probability that an organism can survive and reproduce without the given protein — and functional constraint. They predicted that, given the same functional constraint, proteins with higher dispensability (or lower importance) should evolve faster.

Empirical findings. By Wilson *et al.*'s definition[24], the functional importance of a protein can be measured by the fitness reduction caused by deleting the gene encoding the protein. Thus, one could experimentally test whether the protein evolutionary rate decreases as functional importance increases. However, this test was not feasible until the end of the twentieth century when gene deletion became a routine experiment in several model organisms. The first test performed on a large genetic data set was conducted by Hurst and Smith[25], who classified 175 mouse genes into two groups: essential and non-essential. Essential genes were defined as those that resulted in lethality or infertility when deleted from the mouse genome, whereas deletion of non-essential genes did not. They also measured the nonsynonymous-to-synonymous substitution rate ratio (*d*N/*d*S) (BOX 1) for each mouse gene by comparing with its rat orthologous gene. After removing immunity genes, which are likely to be subject to positive selection, the authors observed no significant difference in *d*N/*d*S between essential and non-essential genes, and they concluded that the functional importance of a protein does not affect its evolutionary rate[25].

By the turn of the century, thousands of genes had been individually deleted from the budding yeast *Saccharomyces cerevisiae* genome. Further studies quantified the growth rates of ~500 of these gene-deletion *S. cerevisiae* strains relative to the wild-type strain[26]. Hirsh and Fraser[27] found, among non-essential genes,

Dispensability
The degree to which an organism can survive and reproduce when a given gene is removed.

Orthologous gene
A gene from a different species that originated by vertical descent from a single gene of the last common ancestor of these species.

a significant but weak negative correlation between the fitness reduction caused by the deletion of a yeast gene and the protein evolutionary rate of the gene.

In the late 1990s, studies were reporting the first genome-wide measurements of gene expression levels using microarray technology[28]. Based on an analysis of these gene expression data, Pal *et al.*[9] reported a strong negative correlation between the expression level of a gene and the evolutionary rate of its protein sequence. They subsequently demonstrated that the correlation reported by Hirsh and Fraser disappeared after controlling for gene expression level[29], suggesting that Hirsh and Fraser's finding was due to covariations between gene expression level with both functional importance and evolutionary rate instead of a causal relationship between functional importance and evolutionary rate.

Zhang and He[30] revisited this hypothesis after additional genomic data from yeast species became available. They found a weak but significant negative correlation between protein evolutionary rate and gene importance defined by the fitness reduction upon gene deletion, with or without controlling for gene expression level. Nonetheless, the partial rank correlation, which is the correlation in rank between two variables after controlling for confounding factors, indicates that only ~1% of the variance in protein evolutionary rate can be explained by the variance in the functional importance of the gene. By contrast, ~25% of the variance in protein evolutionary rate is explainable by the variance in gene expression level. Similar findings were made by Wall *et al.*[31]. Liao *et al.*[32] repeated Hurst and Smith's study[25] with expanded mouse data, reporting a significant negative correlation between gene essentiality and protein evolutionary rate, with or without controlling for gene expression level, although the correlation is again weak. In bacteria, protein functional importance has a statistically significant impact on protein evolutionary rate[33], but the influence of expression level is much greater[34]. In summary, experimental studies in bacteria and eukaryotes demonstrated that the functional importance of a protein only has a weak impact on its evolutionary rate.

Why is the correlation between protein functional importance and evolutionary rate so weak? Wang and Zhang[8] addressed this question from a theoretical perspective, finding that the correlation depends on the distribution of the fitness effects of deleterious mutations (BOX 2). Unfortunately, the lack of empirical data for this distribution prohibited a definitive theoretical prediction on the expected magnitude of the correlation, and this situation has not changed since their study.

Laboratory and natural environments. A caveat of all of the experimental studies of the correlation between protein evolutionary rate and functional importance is that, although evolution occurs in natural environments, functional importance is measured in laboratory conditions. This mismatch is expected to reduce the correlation because it is the functional importance in nature rather than in the laboratory that is predicted to affect the evolutionary rate. Wang and Zhang[8] studied whether the correlation between functional importance and evolutionary rate would be strengthened should functional importance be measured in natural environments. They could not find a strong correlation between evolutionary rate and functional importance measured in any of the 418 laboratory conditions or predicted (for metabolic enzymes) in any of the 10,000 simulated nutritional conditions. Even combinations of the above conditions did not enhance the correlation much. Furthermore, they found no significant difference in evolutionary rates between enzymes that are essential under any nutritional condition and those that are non-essential under any nutritional condition. Taken together, these results strongly suggest that, at least in yeast, the weakness of the correlation is not due to differences in the environment.

Predictive power. Even though these empirical studies found only a weak correlation between functional importance and evolutionary rate, many researchers continue to use sequence conservation to predict functional importance[35] and conclude that the prediction is useful[36]. Wang and Zhang[8] noted that if two yeast proteins were picked at random, there would be a 54% probability that the more slowly evolving protein is functionally more important than the other protein, where functional importance is measured by the fitness effect of gene deletion. This is consistent with the weak correlation reported between these two properties. However, when they ranked all proteins by evolutionary rate and compared two proteins that are separated in rank by more than 95% of all proteins, the probability that the more slowly evolving one is more functionally important than the other becomes 81%. Apparently, the predictive power of the correlation is evident only when proteins with a large difference in evolutionary rate are compared. This also provides an explanation as to why the rate–importance correlation has been successfully used in predicting the functionality of non-coding DNA sequences because most of the reported experimental validations were comparing highly conserved sequences — for example, sequences of at least 200 nucleotides that are identical among humans, mice and rats[36] — with completely unconstrained sequences.

Gene expression level

Gene expression level is a major rate determinant. As mentioned above, Pal *et al.*[9] first reported the unexpected finding that, in yeast, the evolutionary rate of a protein is strongly negatively correlated with its microarray-based mRNA concentration. This negative correlation is often referred to as the E–R anticorrelation, where E stands for gene expression level and R represents evolutionary rate. The E–R anticorrelation exists in all three domains of life[14,34,37], especially when gene expression levels are measured by the more accurate RNA sequencing method instead of the earlier microarray method (FIG. 1). In unicellular organisms, the mRNA concentration of a gene varies across cell cycle stages and environments, but most studies used data collected from the mid-log phase of growth under rich media, which presumably reflect average concentrations across

Effective population size
(Denoted as N_e). A measure of the strength of random genetic drift in a population. The lower the N_e, the stronger the genetic drift. N_e is influenced by the census population size, breeding system and sex ratio, among other factors.

cell cycle stages. In multicellular organisms, mRNA concentration data used are typically from the whole organism or are averaged from several examined tissues. Although the E–R anticorrelation tends to be present regardless of the tissue in which gene expressions are measured, the magnitude of the anticorrelation does vary among tissues[14].

Because of the strong correlation between mRNA and protein concentrations[38,39], the negative correlation between protein concentration and evolutionary rate is also strong[40]. Drummond *et al.*[10] showed that the E–R anticorrelation is weaker when E represents protein concentration than when it stands for mRNA concentration. However, it is unclear whether this disparity is genuine or whether it simply reflects different qualities of proteomic and transcriptomic data. Because proteomic data are available in much fewer species than are transcriptomic data, most studies have used mRNA concentrations in the study of E–R anticorrelation. In this Review, E refers to mRNA concentrations. It is interesting to note that the E–R anticorrelation has also been observed among RNA genes[41,42] (see Supplementary information S1 (box)), and the impact of the expression level of a gene on its evolution extends beyond the sequence level (BOX 3).

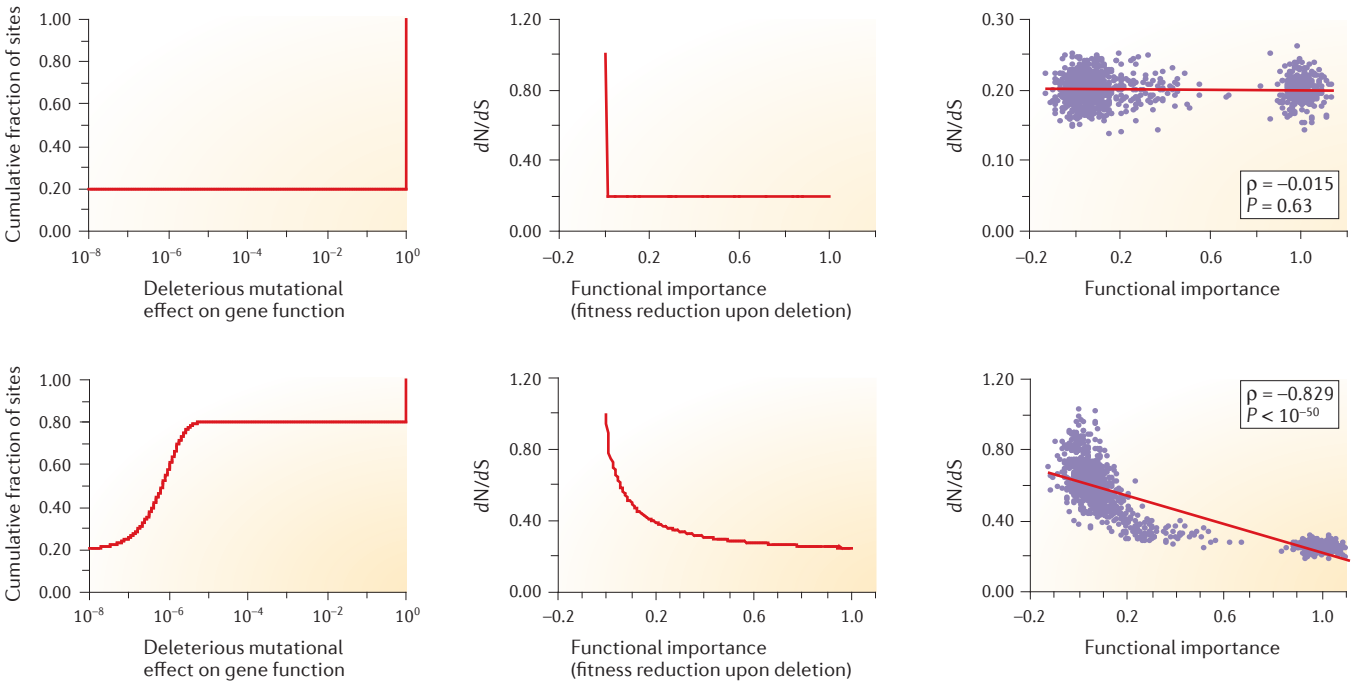
The protein misfolding avoidance hypothesis. What is the mechanism underlying the E–R anticorrelation? This anticorrelation cannot be a by-product of the covariations of protein functional importance with both E and R because the E–R anticorrelation remains strong after controlling for protein functional importance[30,31]. Drummond *et al.*[10] proposed the translational robustness hypothesis to explain the origin of the E–R

Box 2 | Theoretical prediction of the impact of the functional importance of a protein on its evolutionary rate

In qualitative terms, the functional importance of a protein should influence its evolutionary rate, but their quantitative relationship is complex and depends on the effect sizes of mutations. Consider the simplest scenario in which only two kinds of mutations exist: they either completely abolish the function of a gene (with probability α) or do not affect the gene function at all. Let μ be the mutation rate, β be the functional importance (defined as the probability that an organism cannot survive or reproduce without the gene), N be the population size of the organism and N_e be the effective population size. The protein substitution rate (k) in diploid organisms is represented as $k = (1-\alpha)\mu + \alpha(2N\mu)f = \mu[1-\alpha(1-2Nf)]$, where the fixation probability (f) of a new null mutation[6] is as follows.

$$f = \frac{1-e^{\beta N_e/N}}{1-e^{2\beta N_e}}$$

It can be shown[8] that k is a monotonically decreasing function of β. However, f and k are relatively insensitive to β, unless β is on the order of $1/N_e$. Thus, when mutations are null or have no functional effect, a strong negative correlation between functional importance and evolutionary rate is not expected[8]. A hypothetical scenario with $\alpha = 0.8$ is shown in the upper part of the figure, where the left panel depicts the cumulative probability distribution of deleterious functional effects of random mutations, the middle panel shows the theoretical relationship between the functional importance of a gene and the protein evolutionary rate measured by dN/dS (see BOX 1), and the right panel depicts the same relationship for 1,000 genes simulated using functional importance and population size data from the budding yeast when estimation errors were taken into account. Nonetheless, under a different model with the presence of a sizable fraction of deleterious mutations that have functional effects between $1/N_e$ and $100/N_e$, a substantial correlation between functional importance and protein evolutionary rate becomes possible[8]. A hypothetical scenario is shown in the lower part of the figure; in this scenario, 20% of null mutations, 20% of neutral mutations and 60% of slightl y deleterious mutations have functional effects that follow the beta distribution of $\beta(1, 10^6)$. ρ, Spearman's rank correlation coefficient. Figure adapted from REF. 8.



anticorrelation. The key assumption of the hypothesis is that protein misfolding is cytotoxic and reduces fitness[43]. As protein translation is not error-free[44] and translational errors may reduce protein stability and induce protein misfolding, highly expressed genes are under stronger selective pressures than lowly expressed genes to evolve translational robustness that reduces translational error and/or increases protein stability. This requirement for translational robustness constrains sequence evolution. Consequently, more highly expressed proteins evolve more slowly (FIG. 2). A molecular-level simulation confirmed that, under the assumptions of the model, natural

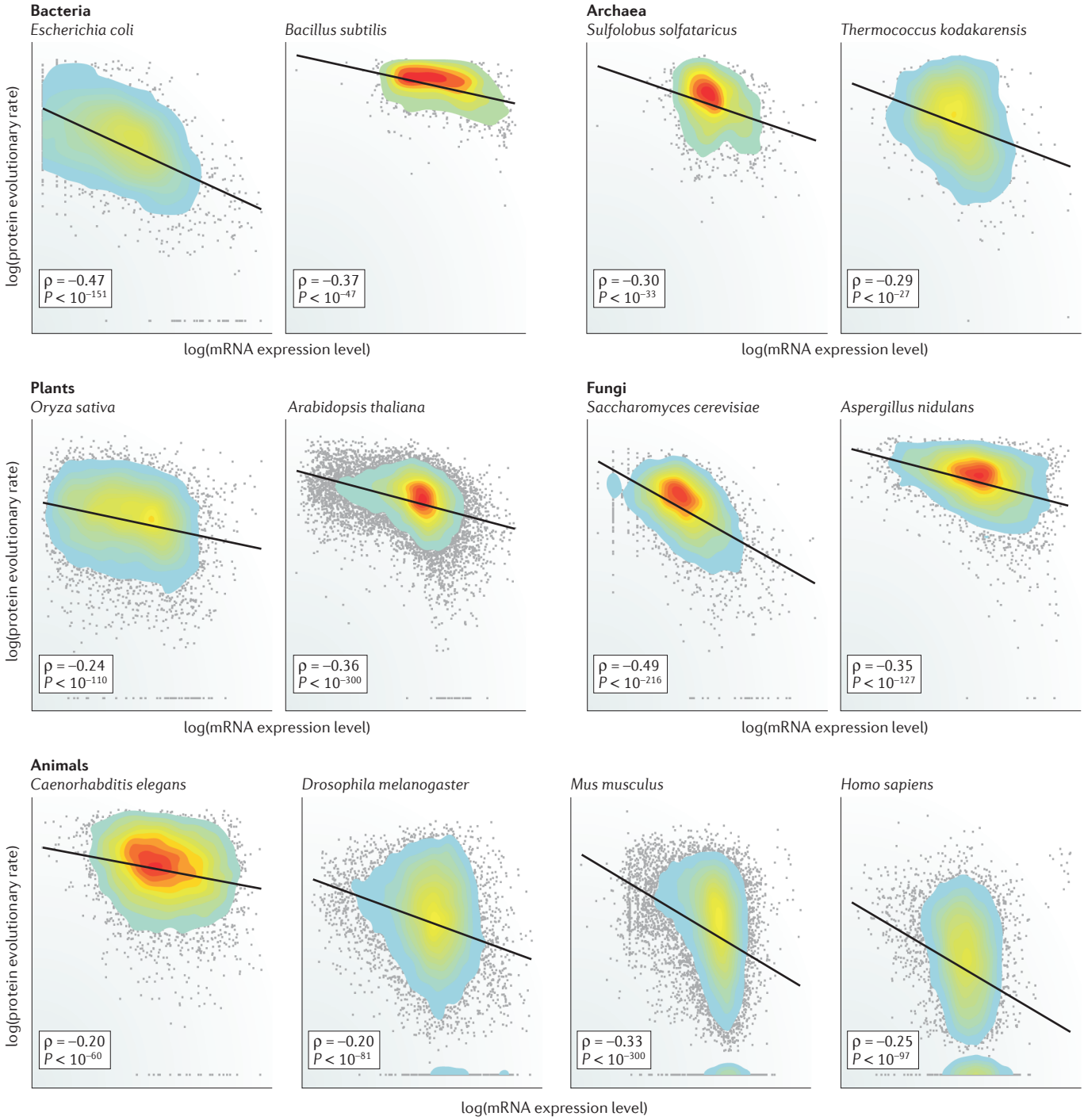


Figure 1 | **The negative correlation between gene expression level and protein evolutionary rate (E–R anticorrelation) exists in all three domains of life.** Protein evolutionary rate is measured by the percentage sequence difference between proteins from a focal species and their orthologous proteins from a closely related species. Each grey dot represents one gene, and the red–light blue gradient represents the density (high to low) of dots such that overplotting is avoided. For each species, the line shows the linear regression, whereas ρ is Spearman's rank correlation coefficient. See Supplementary information S2 (box) for the sources of the data used in making the figure.

Box 3 | **Impact of gene expression level on other aspects of molecular evolution**

In addition to affecting the rate of protein sequence evolution, the expression level of a gene also influences the rates of many other molecular evolutionary events. First, gene expression level affects the mutation rate via two processes: transcription-associated mutagenesis (TAM) and transcription-coupled repair (TCR), which increase and decrease the mutation rate, respectively[91,92]. Recent genomic studies of bacteria, yeast and the human germ line showed that the effect of TAM exceeds that of TCR such that highly expressed genes tend to have high mutation rates[93–96]. In most studies of the E–R anticorrelation, R is estimated by protein divergence, including the effects of both mutation and selection. Because the mutational and selective effects of high expression levels are opposite, the selective effect is expected to be stronger than what the current E–R anticorrelation reveals[93]. Second, transcription is known to induce recombination[97]. Third, high gene expression levels are correlated with a low rate of intertissue expression profile evolution in mammals, although the mechanism remains unclear[98]. Fourth, highly expressed genes are more resistant than lowly expressed genes to gene dosage changes[16]. Fifth, highly expressed genes are less likely than lowly expressed genes to be horizontally transferred in bacteria, probably because the fitness cost of a transfer to the recipient — arising from the energy expenditure in transcription and translation, cytotoxic protein misfolding, reduction in cellular translational efficiency, detrimental protein misinteraction and/or disturbance of the optimal protein concentration or cell physiology — increases with the expression level of the transferred gene, whereas the benefit of the transfer does not increase with the expression level[99].

selection will result in more-stable protein structures, lower translational errors and lower evolutionary rates for more highly expressed proteins[14]. Among various synonymous codons of an amino acid, the preferred codon is believed to be decoded more accurately, and it tends to occupy evolutionarily conserved residues within a gene[45]. In support of the translational robustness hypothesis, Drummond and Wilke[14] found that the favourable use of preferred codons at conserved sites is stronger for the 10% most highly expressed genes than for average genes.

Although Drummond *et al.*'s model encompasses misfolding of correctly translated and mistranslated proteins[14], their studies focused on mistranslation-induced misfolding[10,14]. Yang *et al.*[13] estimated that, depending on the folding stability of a protein, 5–20% of misfolded protein molecules are correctly translated. Their simulation confirmed that the E–R anticorrelation can arise from selection against both translational error-induced misfolding and error-free misfolding, prompting the renaming of the translational robustness hypothesis to the protein misfolding avoidance hypothesis[13] (FIG. 2). The misfolding avoidance hypothesis makes three predictions, each of which has been empirically supported[13]. First, highly expressed proteins were found to have higher folding stabilities than lowly expressed ones[13]. Furthermore, highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins, which are known to be particularly stable[46]. Second, amino acids and codons that increase protein stability were reported to be more prevalent in highly expressed genes than lowly expressed ones[13]. Third, amino acid positions where random mutations would destabilize the protein structure were found to be evolutionarily more conserved than other positions of the same protein[13].

The protein misinteraction avoidance hypothesis. It is well known that amino acid residues located inside a protein structure (that is, core residues) have more central roles than surface residues in protein folding stability[47]. However, surface residues show an E–R anticorrelation as strong as that of core residues, suggesting that selective pressures other than misfolding avoidance might also be present, especially on surface residues[11]. Yang *et al.*[11] showed that the E–R anticorrelation is only moderately weakened by the removal of amino acids that stabilize protein folding and that the impact of this removal on the E–R anticorrelation is smaller when amino acids are removed from protein surfaces than from protein cores. Considering the importance of surface residues in protein–protein interactions, these authors proposed the protein misinteraction avoidance hypothesis[11]. This hypothesis is based on the notion that, even under physiological conditions, proteins may by chance engage in deleterious protein–protein interactions with no physiological function[48–50]. Because the number of misinteracting molecules increases with protein concentration, highly expressed proteins are under a stronger pressure to avoid misinteraction, which constrains their evolution and creates an E–R anticorrelation (FIG. 2). Using computer simulation of a three-dimensional lattice protein model, Yang *et al.*[11] confirmed that selection against deleterious misinteraction results in an E–R anticorrelation. The misinteraction avoidance hypothesis predicts that, compared with lowly expressed proteins, highly expressed proteins disfavour residues that promote misinteraction, exhibit a lower misinteraction probability per molecule and have higher conservation for misinteraction-avoiding residues. These predictions were tested and supported by experimental studies in yeast[11], *Escherichia coli*[51] and humans[51]. Yang *et al.*[11] further predicted that selection against misinteraction should result in translational robustness manifested by reduced mistranslation and reduced misinteraction upon mistranslation, but these predictions have yet to be experimentally tested. As expected, the misinteraction avoidance hypothesis outperforms the misfolding avoidance hypothesis in explaining the E–R anticorrelation for amino acids on protein surfaces[11]. Nevertheless, even together, the two hypotheses seem to be insufficient in providing a full explanation for the anticorrelation because each of them explains only a moderate fraction of the anticorrelation[11].

Protein misfolding
The process by which a protein structure assumes a non-native shape or conformation, which not only diminishes the physiological function of the protein but may also create cytotoxicity.

Preferred codon
A codon that is used more frequently than its synonymous codons in a genome sequence.

Mistranslated proteins
Nascent proteins in which incorrect amino acids have been incorporated during synthesis, which may be caused by incorrect charging of tRNAs by aminoacyl tRNA synthetases or incorrect acceptance of tRNAs by ribosomes.

Protein misinteraction
A non-native interaction between protein molecules that not only reduces the concentrations of freely available protein molecules but may also be toxic.

mRNA folding strength
A measure of the reduction in free energy of a folded mRNA molecule compared to its unfolded form.

The mRNA folding requirement hypothesis. Park *et al.*[12] proposed the mRNA folding requirement hypothesis to explain the E–R anticorrelation. It had been reported that the mRNAs of highly expressed genes have stronger folding (that is, with more negative free energies or higher fractions of paired bases) than those of lowly expressed genes[52]. Park *et al.*[12] showed that this disparity is not a by-product of nucleotide, codon or amino acid compositional differences among genes of different expression levels but results from intensified selection for strong mRNA folding in highly expressed genes. If high-concentration mRNAs have been selected for strong folding, then a random mutation is more likely to reduce mRNA folding and be harmful when occurring in a highly expressed gene than in a lowly expressed gene. Consequently, the higher the gene expression level, the lower the substitution rate, creating an E–R anticorrelation (FIG. 2). In support of this hypothesis, Park *et al.*[12] detected a strong negative correlation between mRNA folding strength and protein evolutionary rate both before and after controlling for gene expression level.

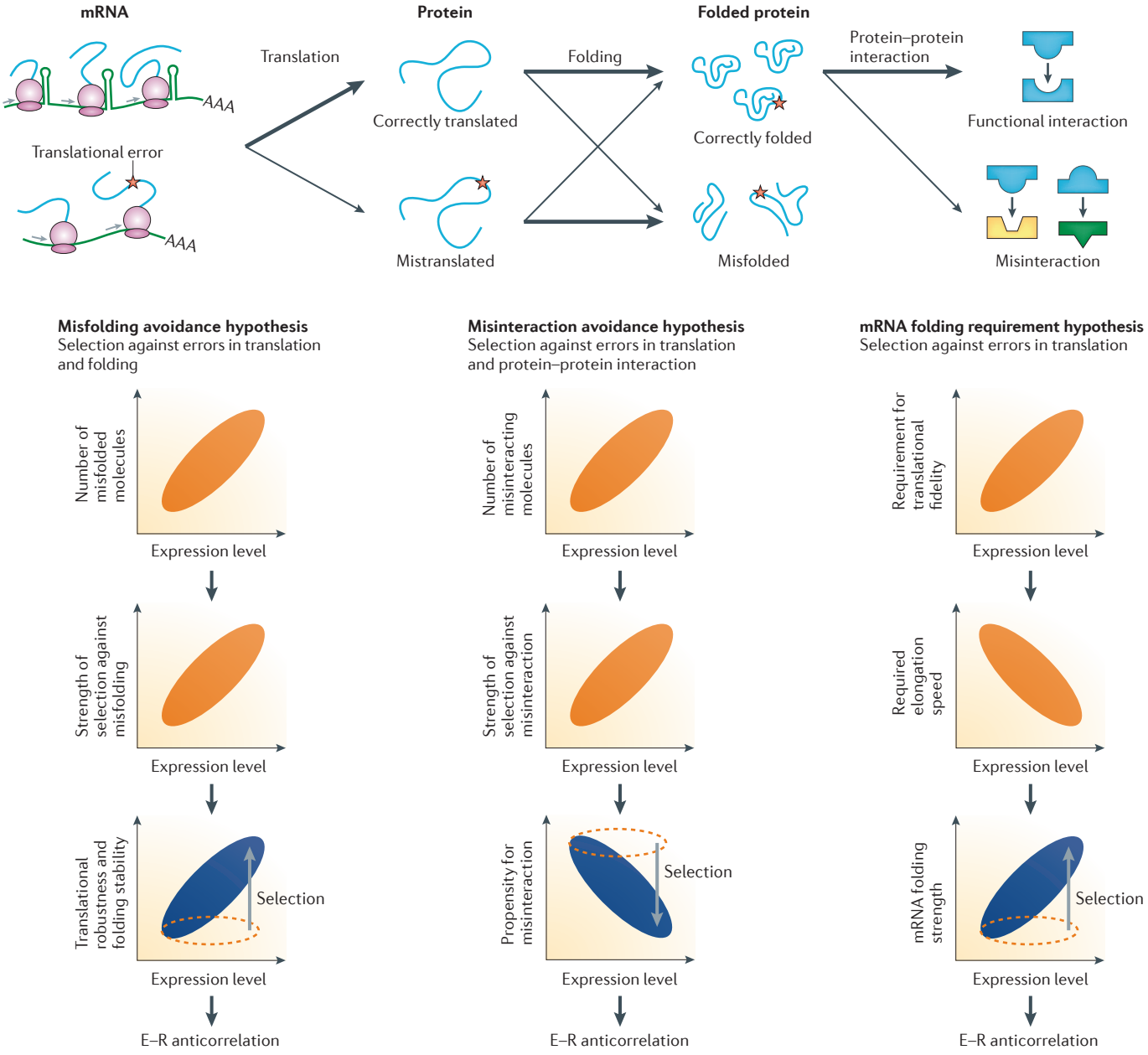


Figure 2 | **Natural selection against errors in protein translation, folding and interaction can explain the E–R anticorrelation.** The upper part of the figure shows key molecular processes in the production and functioning of proteins, as well as the types of error generated in these processes. The lower part of the figure shows expected relationships between the expression level of a protein and the properties of the protein in relation to the various errors mentioned, providing rationales for the hypotheses that natural selection against molecular errors generates the E–R anticorrelation. The orange and dark blue ovals show relationships between expression level and various gene properties before and after selection, respectively.

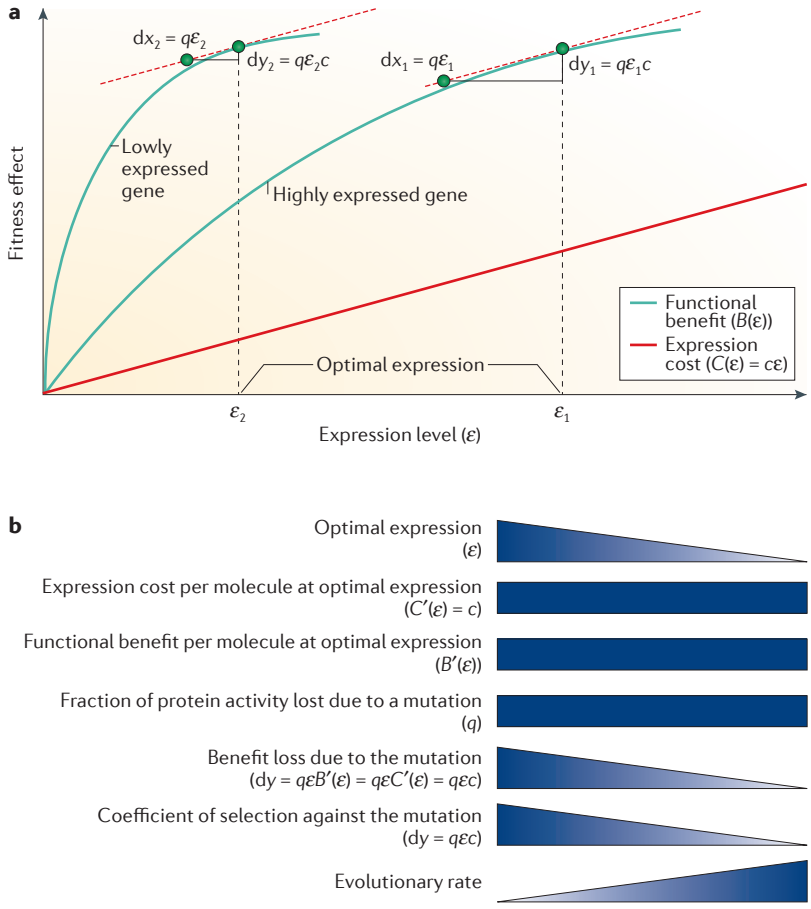


Figure 3 | **The expression cost hypothesis of the E–R anticorrelation.** **a** | The functional benefit of the expression of a gene and the cost of expression for each gene are plotted. The graph shows that the functional benefit and the cost of expressing an extra molecule at the optimal expression level are equal. A mutation that reduces protein activity by a fraction q imposes a bigger loss of benefit for the highly expressed gene (dy_1) than the lowly expressed gene (dy_2). **b** | The expected trends of various gene properties as the optimal expression level (shown by the top bar) decreases. The height of a symbol represents the quantity concerned. ε represents the optimal protein abundance, B and C denote the benefit and cost of the expression, respectively, and c represents expression cost per molecule.

However, why would more highly expressed genes be under selection for stronger mRNA folding? It was recently discovered that strong mRNA folding at the leading edge of an elongating ribosome slows decoding at the ribosome A site and increases translational accuracy due to a trade-off[15]. Fast elongation is beneficial in alleviating ribosome sequestration when ribosomes are in shortage during rapid cell growth, but fast elongation is also costly because it compromises translational fidelity, which would waste material and energy for protein synthesis and induce toxic protein misfolding and misinteraction. One theoretical study modelled the cost and benefit of fast elongation and found that the optimal ribosomal elongation speed decreases as the expression level of the protein increases[15]. In short, the requirement for stronger mRNA folding of more highly expressed genes is thought to be attributable to the demand for translational accuracy, but whether this is the sole reason is unknown.

The expression cost hypothesis. Gout *et al.*[16] and Cherry[17] independently proposed the expression cost hypothesis to explain the E–R anticorrelation (FIG. 3). This hypothesis is based on two assumptions. First, protein synthesis has associated cost (C) and benefit (B) that are both increasing functions of protein abundance. Second, the optimal protein abundance (ε) is reached when the rate of increase in C with protein abundance equals that of B; that is, if one more protein molecule is synthesized at the optimal expression level measured by protein abundance, the extra benefit should equal the extra cost: $B'(\varepsilon) = C'(\varepsilon)$, where the prime symbol denotes the derivative. Hence, a mutation that decreases the protein activity by a small fraction q, having a functional effect equivalent to the loss of $q\varepsilon$ molecules, will reduce the fitness by $q\varepsilon B'(\varepsilon) = q\varepsilon C'(\varepsilon)$. Thus, if $C'(\varepsilon)$ is constant among genes, the higher the value of ε, the stronger the selection against the deleterious mutation with a given q, leading to an E–R anticorrelation. Under the expression cost hypothesis, the E–R anticorrelation results from selection against mutations disrupting protein physiological functions, unlike all other hypotheses proposing that the anticorrelation arises from selection against mutations enhancing protein toxicity. Nonetheless, the functional importance of a protein, measured in this model by $B(\varepsilon) - C(\varepsilon)$, may be different for two proteins with the same expression level, whereas proteins with the same functional importance may have different expression levels. The expression cost hypothesis predicts that the strength of selection against deleterious mutations is determined by the expression level rather than by functional importance.

However, the validity of the elegant expression cost hypothesis in explaining the E–R anticorrelation has not been extensively investigated empirically. One piece of evidence used to support this hypothesis is that deleting an allele of a highly expressed gene from a diploid yeast tends to cause more harm than deleting an allele of a lowly expressed gene[16], but it is unclear whether this phenomenon results from the expression cost hypothesis or simply a by-product of the correlation between functional importance and expression level that is unrelated to the expression cost hypothesis. Furthermore, it is not a precise test of the expression cost hypothesis because the test is conducted for mutations with $q = 0.5$, whereas the hypothesis requires $q << 1$.

What constituents are included in the cost of protein expression is another crucial question. It certainly should include the material and energy costs (that is, synthetic costs) of transcription and translation, which are proportional to the product of protein length and expression level. Intriguingly, if the expression cost is entirely due to the synthetic cost, proteins of different lengths should have different expression costs per molecule, and the expression cost hypothesis would no longer predict slower evolution of more highly expressed proteins. What it would predict is slower evolution of more highly expressed proteins upon controlling for protein length and slower evolution of longer proteins upon controlling for expression level. Based on our analysis of yeast data, the former prediction is

supported but the latter is not, suggesting that the synthetic cost is at most a minor component of the protein expression cost. Presumably, the expression cost also includes the deleterious effects of protein mistranslation, misfolding and misinteraction. Several studies showed reduced per-molecule cost of mistranslation, misfolding and misinteraction for highly expressed proteins compared with lowly expressed ones[11,13–15]. There is also evidence that highly expressed proteins tend to use amino acids with relatively low synthetic costs[53]. In other words, $C'(\varepsilon)$ becomes smaller as ε increases. In the extreme case, $\varepsilon C'(\varepsilon)$ may become similar among genes with different ε. Consequently, larger ε no longer results in stronger purifying selection, leading to the collapse of the hypothesis. The expression cost hypothesis is probably correct to some extent, but its importance, relative to the other hypotheses, in explaining the E–R anticorrelation requires further investigation.

Correlates of protein evolutionary rate

In addition to the factors discussed above, numerous other correlates of the protein evolutionary rate have been reported (TABLE 1). Of particular interest is the effect of the fusion of a pair of domains in multidomain proteins on the domain-specific evolutionary rates[54]. Wolf *et al.*[54] discovered that domains with substantially different evolutionary rates in separate proteins retain these domain-specific rates to some extent even within the context of multidomain proteins. This suggests the importance of domain-specific features in determining the protein evolutionary rate, but it is unclear what these features are. They could be constraints arising from domain-specific functions but could also be domain-specific probabilities of protein misfolding and/or misinteraction.

Many reported rate determinants in TABLE 1 have small effects, although a few seem to show moderate impacts. Nevertheless, the mechanisms behind their direct or indirect impacts are often unknown. For instance, the number of types of microRNA targeting a mammalian gene is the best predictor of the protein evolutionary rate of the gene[55], and its impact goes beyond those of gene expression level[55] and 3′ untranslated region length[56]. One suggested mechanism is that the number of microRNA type reflects the pleiotropic level of the target gene[55], which is known to constrain protein evolution[57]. However, it is unclear why this number measures the pleiotropic level, which by definition is the number of functions of the gene, and exactly how pleiotropy constrains protein evolution. Furthermore, if pleiotropy is a primary rate determinant, it is puzzling why the number of protein interaction partners of a protein, which is presumably a reliable measure of pleiotropy, has only a minor effect on protein evolutionary rate[58,59]. Understanding why these and other factors do or do not affect protein evolutionary rates will be an important task for the future.

All of the factors discussed so far affect the intensity of purifying selection, which prevents the fixation of deleterious mutations. In theory, the rate of protein evolution is also affected by positive selection, which promotes the fixation of beneficial mutations. However, because the vast majority of mutations are deleterious, the impact of purifying selection far exceeds that of positive selection in the evolution of almost all proteins. The prominent impact of positive selection on the rate of protein evolution is evident in only a small fraction of proteins, mainly those subject to recurrent positive selection that is typically related to host–pathogen interactions[60] or intersexual interactions[61]. For this reason, factors pertaining to positive selection are usually ignored in the search for correlates of protein evolutionary rate, but whether this negligence affects our understanding of the rate determinants remains to be studied.

Implications for biology and medicine

As reviewed here, the functional importance of a protein is not a major contributor to functional constraint in the evolution of the protein. Based on our current understanding of the major correlates of the rate of protein sequence evolution and their underlying causes, important components of the functional constraint include propensities for several types of molecular and cellular errors, such as mistranslation, misfolding and misinteraction. As a result, the word 'functional' in 'functional constraint' should not be interpreted exclusively or even primarily as relating to physiological function but should also include toxicity (or negative function). In other words, mutations can be unacceptable owing to the disruption of a physiological function or the creation of toxicity. This is a substantially expanded understanding of protein evolution from the standard explanation that has dominated evolutionary biology for nearly 50 years.

There are several important biological implications of this new understanding of protein evolution. First, the evolutionary rate of a protein only reflects to a very small extent the importance of the protein's physiological function. Inference of the relative importance of proteins from their evolutionary rates is expected to be unreliable, unless there is a large difference in their evolutionary rates. Second, a lower-than-neutral rate of sequence evolution suggests that the sequence is constrained, but the reason could be the existence of a physiological function or a propensity for one or more toxic cellular and molecular errors. For instance, translational stop codon readthrough has been reported for hundreds of fruitfly genes, and the average evolutionary rate of the translated regions downstream of the stop codons is slightly lower than that of neutral sequences[62]. This observation is not a proof that the post-stop-codon regions have physiological functions because the observation could be due to toxicity avoidance constraints. Third, studies in the past decade have revealed many stochastic errors and noises in cellular and molecular processes such as gene expression[63] and pre-mRNA splicing[64]. This is hardly surprising because these processes require biochemical reactions, which are stochastic in nature. The biological importance of these errors and noises is only starting to be understood[65–69], and the discovery of their dominant roles in shaping protein evolution points to a potential of their involvement in all aspects of cellular and molecular biology as well as evolution.

Pleiotropic
Pertaining to pleiotropy: the phenomenon whereby one gene or one mutation affects multiple traits.

Designability
The number of protein sequences that adopt a protein structure.

Protein conformational diversity
The degree of structural variations of various native states of a protein.

Table 1 | **Correlates of the protein evolutionary rate**

Correlates	Properties of faster evolving proteins	Organisms	Refs
Gene expression level	Lower expressions	Bacteria, archaea and eukaryotes	9,14,34, 37
Functional importance	Lower importance and higher dispensability	Yeast and mammals	27,30–32
Expression breadth among tissues	Lower expression breadth and higher tissue specificity	Mammals	32,100
Expression timing in development	Expression in late embryogenesis and adulthood	Zebrafish	101
Promoter and gene body methylation	Higher levels of promoter methylation but lower gene body methylation levels	Mammals	102
Chaperone targeting	Higher levels of chaperone targeting	Bacteria and eukaryotes	103,104
Protein subcellular localization	Higher tendency to be extracellular	Yeast and mammals	105
Codon usage bias	Weaker codon usage bias	Bacteria and eukaryotes	14
Distance from the origin of replication	Larger distance	Bacteria and archaea	106, 107,108
Pleiotropy	Lower pleiotropy	Eukaryotes	57
Protein–protein interaction network properties	Lower connectivity, closeness and betweenness	Eukaryotes	58,109 ,110
Metabolic network property	Lower flux and connectivity	Yeast	111
Regulatory network properties	Higher centrality	Yeast	112
Targeting by microRNAs	Fewer types of targeting microRNA	Mammals	55
Gene compactness	Shorter introns and untranslated regions	Mammals	32
Protein length	Longer proteins	Yeast and mammals	113,114
mRNA folding	Weaker mRNA folding	Bacteria and eukaryotes	12
GC content	Lower GC content	Mammals	113
Domain structure	Lower density of domains	Animals and plants	54
Protein disordered regions	More-disordered regions	Bacteria and eukaryotes	115
Protein structural designability	Higher inter-residue contact density and higher fraction of buried sites	Yeast	114
Protein conformational diversity	Lower conformational diversity	Mammals	116

The new understanding of protein evolution also has medical implications. First, the notion that a large fraction of unacceptable mutations are not loss-of-function mutations but gain-of-toxicity mutations provides new insights into the mechanistic basis of certain genetic diseases. For example, misfolding of proteins is known to cause various diseases[70]. Computational analysis has suggested that up to 80% of disease-causing missense mutations reduce protein structural stability, which would increase the misfolding probability[71]. Similarly, hydrophilic-to-hydrophobic mutations on the surface of a highly expressed protein could induce deleterious protein misinteraction, as seen in some mutants of the tumour suppressor gene *TP53* (REF. 72). Furthermore, overexpression of a promiscuous protein that has low expression levels under normal conditions could induce disease-causing protein misinteraction, as demonstrated in cancers[49]. Second, the prediction of disease-causing mutations is of substantial medical importance and is a rapidly growing field[73,74]. The new understanding of critical factors constraining protein evolution allows better predictions of potentially harmful mutations and the associated mechanisms. Third, although natural selection has reduced the rates of several molecular and cellular errors discussed in this Review, somatic mutations could bring them back to high levels. Whether increased error rates caused by somatic mutations are partially responsible for ageing[75,76], cancer[77] and other diseases[78] is worth systematic investigation.

Conclusions and future studies

Studies of the rate of protein evolution began with the field of molecular evolution in the 1960s and have been recently renewed by the wide availability of genomic data. Although these recent studies have uncovered unsuspected forces in protein evolution, several questions remain for future studies. First, although the expression cost hypothesis of the E–R anticorrelation is theoretically attractive, it still lacks definitive empirical evidence, and the main components of the expression cost have not been specified. Second, the interdependency, relative contributions and combined

explanatory power of the multiple identified causes of the E–R anticorrelation are unclear, and it is unknown whether additional causes exist. Third, apart from mistranslation, protein misfolding and protein misinteraction, other cellular and molecular errors — such as transcriptional error, splicing error, RNA editing error, translation initiation from upstream start codons and translational stop codon readthrough — have not been investigated for their potential effects on the protein evolutionary rate. Fourth, the distribution of the fitness effects of mutations in a gene plays an important part in determining the rate of protein sequence evolution (BOX 1), but the details of this distribution have not been elucidated empirically. Recent studies using high-throughput next-generation DNA sequencing methods are making progress in characterizing this distribution[79,80]. Nevertheless, the functional basis of the fitness distribution is more difficult to identify and may, for example, involve weakening of the physiological function of a protein and enhancement of its cytotoxicity. Fifth, because cellular errors may by chance create new protein variants, it would be interesting to study whether these errors have important roles in the origin of new protein functions and adaptation[81]. Sixth, the mechanisms underlying the correlations between the protein evolutionary rate and many of the factors listed in TABLE 1 are unknown, and an integrative approach is required for understanding the interdependencies among these factors[82–84]. Seventh, how much of what we have learned about the evolutionary rate of proteins apply to the evolutionary rate of non-coding RNAs is a largely unexplored area (see Supplementary information S1 (box)). Eighth, the evolutionary rate of a particular protein can change during the course of evolution, but major factors underlying such changes remain largely unknown[85]. Last, this Review is focused on the variation of evolutionary rate among different proteins rather than among different regions of a protein. Although the latter has been extensively studied[86,87], the connection between the two variations is not well understood[88,89]. By answering these major unsolved questions, the study of protein evolutionary rate holds promise to offer further insights into the mechanisms of evolution and disease.

1. Zuckerkandl, E. & Pauling, L. in *Horizons in Biochemistry* (eds Kasha, M. & Pullman, B.) 189–225 (Academic Press, 1962).
2. Zuckerkandl, E. & Pauling, L. in *Evolving Genes and Proteins* (eds Bryson, V. & Vogel, H. J.) 97–166 (Academic Press, 1965).
3. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
4. Kumar, S. Molecular clocks: four decades of evolution. *Nat. Rev. Genet.* **6**, 654–662 (2005).
5. Takahata, N. Molecular clock: an anti-neo-Darwinian legacy. *Genetics* **176**, 1–6 (2007).
6. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, 1983).
7. Li, W. *Molecular Evolution* (Sinauer, 1997).
8. Wang, Z. & Zhang, J. Why is the correlation between gene importance and gene evolutionary rate so weak? *PLoS Genet.* **5**, e1000329 (2009).
9. Pal, C., Papp, B. & Hurst, L. D. Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927–931 (2001).
This is the first report of the E–R anticorrelation.
10. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA* **102**, 14338–14343 (2005).
This paper proposes the translational robustness hypothesis of the E–R anticorrelation.
11. Yang, J. R., Liao, B. Y., Zhuang, S. M. & Zhang, J. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc. Natl Acad. Sci. USA* **109**, E831–E840 (2012).
This paper proposes the protein misinteraction hypothesis of the E–R anticorrelation.
12. Park, C., Chen, X., Yang, J. R. & Zhang, J. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA* **110**, E678–E686 (2013).
This paper proposes the mRNA folding requirement hypothesis of the E–R anticorrelation.
13. Yang, J. R., Zhuang, S. M. & Zhang, J. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol. Syst. Biol.* **6**, 421 (2010).
14. Drummond, D. A. & Wilke, C. O. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352 (2008).
15. Yang, J. R., Chen, X. & Zhang, J. Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS Biol.* **12**, e1001910 (2014).
This paper explains the underlying cause of the mRNA folding requirement that partially accounts for the E–R anticorrelation.
16. Gout, J. F., Kahn, D. & Duret, L. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet.* **6**, e1000944 (2010).
17. Cherry, J. L. Expression level, evolutionary rate, and the cost of expression. *Genome Biol. Evol.* **2**, 757–769 (2010).
References 16 and 17 independently propose the expression cost hypothesis of the E–R anticorrelation.
18. Yang, Z. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* **11**, 367–372 (1996).
19. Hedges, S. B., Dudley, J. & Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971–2972 (2006).
20. King, J. L. & Jukes, T. H. Non-Darwinian evolution. *Science* **164**, 788–798 (1969).
21. Zhang, J. in *Evolution Since Darwin: The First 150 Years* (eds Bell, M. A. *et al.*) 87–118 (Sinauer, 2010).
22. Karp, G. *Cell and Molecular Biology* (John Wiley & Sons, 2008).
23. Kimura, M. & Ohta, T. On some principles governing molecular evolution. *Proc. Natl Acad. Sci. USA* **71**, 2848–2852 (1974).
This paper proposes the role of protein functional importance and functional constraint in determining the rate of protein sequence evolution.
24. Wilson, A. C., Carlson, S. S. & White, T. J. Biochemical evolution. *Annu. Rev. Biochem.* **46**, 573–639 (1977).
25. Hurst, L. D. & Smith, N. G. Do essential genes evolve slowly? *Curr. Biol.* **9**, 747–750 (1999).
This study was the first to test the relationship between protein functional importance and evolutionary rate based on a fairly large genomic data set.
26. Winzeler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
27. Hirsh, A. E. & Fraser, H. B. Protein dispensability and rate of evolution. *Nature* **411**, 1046–1049 (2001).
28. Holstege, F. C. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717–728 (1998).
29. Pal, C., Papp, B. & Hurst, L. D. Genomic function: rate of evolution and gene dispensability. *Nature* **421**, 496–497 (2003).
30. Zhang, J. & He, X. Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol. Biol. Evol.* **22**, 1147–1155 (2005).
31. Wall, D. P. *et al.* Functional genomic analysis of the rates of protein evolution. *Proc. Natl Acad. Sci. USA* **102**, 5483–5488 (2005).
32. Liao, B. Y., Scott, N. M. & Zhang, J. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Mol. Biol. Evol.* **23**, 2072–2080 (2006).
33. Jordan, I. K., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* **12**, 962–968 (2002).
34. Rocha, E. P. & Danchin, A. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol. Biol. Evol.* **21**, 108–116 (2004).
35. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
36. Pennacchio, L. A. *et al. In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
37. Krylov, D. M., Wolf, Y. I., Rogozin, I. B. & Koonin, E. V. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* **13**, 2229–2235 (2003).
38. Greenbaum, D., Colangelo, C., Williams, K. & Gerstein, M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.* **4**, 117 (2003).
39. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232 (2012).
40. Drummond, D. A., Raval, A. & Wilke, C. O. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* **23**, 327–337 (2006).
41. Shen, Y. *et al.* Testing hypotheses on the rate of molecular evolution in relation to gene expression using microRNAs. *Proc. Natl Acad. Sci. USA* **108**, 15942–15947 (2011).
42. Managadze, D., Rogozin, I. B., Chernikova, D., Shabalina, S. A. & Koonin, E. V. Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol. Evol.* **3**, 1390–1404 (2011).
43. Geiler-Samerotte, K. A. *et al.* Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc. Natl Acad. Sci. USA* **108**, 680–685 (2011).
44. Drummond, D. A. & Wilke, C. O. The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.* **10**, 715–724 (2009).
45. Akashi, H. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136**, 927–935 (1994).

46. Cherry, J. L. Highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins. *Mol. Biol. Evol.* **27**, 735–741 (2010).
47. Chakravarty, S. & Varadarajan, R. Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure* **7**, 723–732 (1999).
48. Stambolsky, P. *et al.* Modulation of the vitamin D3 response by cancer-associated mutant p53. *Cancer Cell* **17**, 273–285 (2010).
49. Vavouri, T., Semple, J. I., Garcia-Verdugo, R. & Lehner, B. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell* **138**, 198–208 (2009).
50. Zhang, J., Maslov, S. & Shakhnovich, E. I. Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size. *Mol. Syst. Biol.* **4**, 210 (2008).
51. Levy, E. D., De, S. & Teichmann, S. A. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proc. Natl Acad. Sci. USA* **109**, 20461–20466 (2012).
52. Zur, H. & Tuller, T. Strong association between mRNA folding strength and protein abundance in *S. cerevisiae*. *EMBO Rep.* **13**, 272–277 (2012).
53. Akashi, H. & Gojobori, T. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc. Natl Acad. Sci. USA* **99**, 3695–3700 (2002).
54. Wolf, M. Y., Wolf, Y. I. & Koonin, E. V. Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution. *Biol. Direct* **3**, 40 (2008).
55. Chen, S. C., Chuang, T. J. & Li, W. H. The relationships among microRNA regulation, intrinsically disordered regions, and other indicators of protein evolutionary rate. *Mol. Biol. Evol.* **28**, 2513–2520 (2011).
56. Cheng, C., Bhardwaj, N. & Gerstein, M. The relationship between the evolution of microRNA targets and the length of their UTRs. *BMC Genomics* **10**, 431 (2009).
57. He, X. & Zhang, J. Toward a molecular understanding of pleiotropy. *Genetics* **173**, 1885–1891 (2006).
58. Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. & Feldman, M. W. Evolutionary rate in the protein interaction network. *Science* **296**, 750–752 (2002).
59. Bloom, J. D. & Adami, C. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein–protein interactions data sets. *BMC Evol. Biol.* **3**, 21 (2003).
60. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170 (1988).
61. Lee, Y. H., Ota, T. & Vacquier, V. D. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Mol. Biol. Evol.* **12**, 231–238 (1995).
62. Dunn, J. G., Foo, C. K., Belletier, N. G., Gavis, E. R. & Weissman, J. S. Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster*. *eLife* **2**, e01179 (2013).
63. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–226 (2008).
64. Marinov, G. K. *et al.* From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res.* **24**, 496–510 (2014).
65. Zhang, Z., Qian, W. & Zhang, J. Positive selection for elevated gene expression noise in yeast. *Mol. Syst. Biol.* **5**, 299 (2009).
66. Wang, Z. & Zhang, J. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proc. Natl Acad. Sci. USA* **108**, E67–E76 (2011).
67. Warnecke, T. & Hurst, L. D. Error prevention and mitigation as forces in the evolution of genes and genomes. *Nat. Rev. Genet.* **12**, 875–881 (2011).
68. Eldar, A. & Elowitz, M. B. Functional roles for noise in genetic circuits. *Nature* **467**, 167–173 (2010).
69. Xu, G. & Zhang, J. Human coding RNA editing is generally nonadaptive. *Proc. Natl Acad. Sci. USA* **111**, 3769–3774 (2014).
70. Gregersen, N., Bross, P., Vang, S. & Christensen, J. H. Protein misfolding and human disease. *Annu. Rev. Genom. Hum. Genet.* **7**, 103–124 (2006).
71. Wang, Z. & Moult, J. SNPs, protein structure, and disease. *Hum. Mutat.* **17**, 263–270 (2001).
72. Oren, M. & Rotter, V. Mutant p53 gain-of-function in cancer. *Cold Spring Harb. Perspect Biol* **2**, a001107 (2010).
73. Wu, J., Li, Y. & Jiang, R. Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies. *PLoS Genet.* **10**, e1004237 (2014).
74. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12**, 628–640 (2011).
75. Orgel, L. E. The maintenance of the accuracy of protein synthesis and its relevance to ageing. *Proc. Natl Acad. Sci. USA* **49**, 517–521 (1963).
76. Silva, R. M. *et al.* The yeast *PNC1* longevity gene is up-regulated by mRNA mistranslation. *PLoS ONE* **4**, e5212 (2009).
77. Pandolfi, P. P. Aberrant mRNA translation in cancer pathogenesis: an old concept revisited comes finally of age. *Oncogene* **23**, 3134–3137 (2004).
78. Frank, S. A. Somatic mosaicism and disease. *Curr. Biol.* **24**, R577–R581 (2014).
79. Stiffler, M. A., Hekstra, D. R. & Ranganathan, R. Evolvability as a function of purifying selection in TEM-1 β-lactamase. *Cell* **160**, 882–892 (2015).
80. Podgornaia, A. I. & Laub, M. T. Pervasive degeneracy and epistasis in a protein–protein interface. *Science* **347**, 673–677 (2015).
81. Whitehead, D. J., Wilke, C. O., Vernazobres, D. & Bornberg-Bauer, E. The look-ahead effect of phenotypic mutations. *Biol. Direct* **3**, 18 (2008).
82. Pal, C., Papp, B. & Lercher, M. J. An integrated view of protein evolution. *Nat. Rev. Genet.* **7**, 337–348 (2006).
83. Wolf, Y. I., Carmel, L. & Koonin, E. V. Unifying measures of gene function and evolution. *Proc. Biol. Sci.* **273**, 1507–1515 (2006).
84. Xia, Y., Franzosa, E. A. & Gerstein, M. B. Integrated assessment of genomic correlates of protein evolutionary rate. *PLoS Comput. Biol.* **5**, e1000413 (2009).
85. Du, X., Lipman, D. J. & Cherry, J. L. Why does a protein's evolutionary rate vary over time? *Genome Biol. Evol.* **5**, 494–503 (2013).
86. Franzosa, E. A. & Xia, Y. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.* **26**, 2387–2395 (2009).
87. Yeh, S. W. *et al.* Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure. *Mol. Biol. Evol.* **31**, 135–139 (2014).
88. Zhang, J. & Gu, X. Correlation between the substitution rate and rate variation among sites in protein evolution. *Genetics* **149**, 1615–1625 (1998).
89. Chen, F. C., Liao, B. Y., Pan, C. L., Lin, H. Y. & Chang, A. Y. Assessing determinants of exonic evolutionary rates in mammals. *Mol. Biol. Evol.* **29**, 3121–3129 (2012).
90. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, 2000).
91. Kim, N. & Jinks-Robertson, S. Transcription as a source of genome instability. *Nat. Rev. Genet.* **13**, 204–214 (2012).
92. Hanawalt, P. C. & Spivak, G. Transcription-coupled DNA repair: two decades of progress and surprises. *Nat. Rev. Mol. Cell Biol.* **9**, 958–970 (2008).
93. Park, C., Qian, W. & Zhang, J. Genomic evidence for elevated mutation rates in highly expressed genes. *EMBO Rep.* **13**, 1123–1129 (2012).
94. Chen, X. & Zhang, J. No gene-specific optimization of mutation rate in *Escherichia coli*. *Mol. Biol. Evol.* **30**, 1559–1562 (2013).
95. Chen, X. & Zhang, J. Yeast mutation accumulation experiment supports elevated mutation rates at highly transcribed sites. *Proc. Natl Acad. Sci. USA* **111**, E4062 (2014).
96. Lind, P. A. & Andersson, D. I. Whole-genome mutational biases in bacteria. *Proc. Natl Acad. Sci. USA* **105**, 17878–17883 (2008).
97. Gottipati, P. & Helleday, T. Transcription-associated recombination in eukaryotes: link between transcription, replication and recombination. *Mutagenesis* **24**, 203–210 (2009).
98. Liao, B. Y. & Zhang, J. Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Mol. Biol. Evol.* **23**, 1119–1128 (2006).
99. Park, C. & Zhang, J. High expression hampers horizontal gene transfer. *Genome Biol. Evol.* **4**, 523–532 (2012).
100. Zhang, L. & Li, W. H. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol. Biol. Evol.* **21**, 236–239 (2004).
101. Piasecka, B., Lichocki, P., Moretti, S., Bergmann, S. & Robinson-Rechavi, M. The hourglass and the early conservation models — co-existing patterns of developmental constraints in vertebrates. *PLoS Genet.* **9**, e1003476 (2013).
102. Chuang, T. J. & Chiang, T. W. Impacts of pretranscriptional DNA methylation, transcriptional transcription factor, and posttranscriptional microRNA regulations on protein evolutionary rate. *Genome Biol. Evol.* **6**, 1530–1541 (2014).
103. Taipale, M. *et al.* Quantitative analysis of HSP90-client interactions reveals principles of substrate recognition. *Cell* **150**, 987–1001 (2012).
104. Bogumil, D. & Dagan, T. Chaperonin-dependent accelerated substitution rates in prokaryotes. *Genome Biol. Evol.* **2**, 602–608 (2010).
105. Liao, B. Y., Weng, M. P. & Zhang, J. Impact of extracellularity on the evolutionary rate of mammalian proteins. *Genome Biol. Evol.* **2**, 39–43 (2010).
106. Ran, W., Kristensen, D. M. & Koonin, E. V. Couping between protein level selection and codon usage optimization in the evolution of bacteria and archaea. *mBio* **5**, e00956-14 (2014).
107. Sharp, P. M., Shields, D. C., Wolfe, K. H. & Li, W. H. Chromosomal location and evolutionary rate variation in enterobacterial genes. *Science* **246**, 808–810 (1989).
108. Flynn, K. M., Vohr, S. H., Hatcher, P. J. & Cooper, V. S. Evolutionary rates and gene dispensability associate with replication timing in the archaeon *Sulfolobus islandicus*. *Genome Biol. Evol.* **2**, 859–869 (2010).
109. Hahn, M. W. & Kern, A. D. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* **22**, 803–806 (2005).
110. Kim, P. M., Lu, L. J., Xia, Y. & Gerstein, M. B. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314**, 1938–1941 (2006).
111. Vitkup, D., Kharchenko, P. & Wagner, A. Influence of metabolic network structure and function on enzyme evolution. *Genome Biol.* **7**, R39 (2006).
112. Jovelin, R. & Phillips, P. C. Evolutionary rates and centrality in the yeast gene regulatory network. *Genome Biol.* **10**, R35 (2009).
113. Kryuchkova, N. & Robinson-Rechavi, M. Determinants of protein evolutionary rates in light of ENCODE functional genomics. *BMC Bioinformatics* **15** (Suppl. 3), A9 (2014).
114. Bloom, J. D., Drummond, D. A., Arnold, F. H. & Wilke, C. O. Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.* **23**, 1751–1761 (2006).
115. Brown, C. J. *et al.* Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* **55**, 104–110 (2002).
116. Javier Zea, D., Miguel Monzon, A., Fornasari, M. S., Marino-Buslje, C. & Parisi, G. Protein conformational diversity correlates with evolutionary rate. *Mol. Biol. Evol.* **30**, 1500–1503 (2013).

Acknowledgements
The authors thank X. Chen, W.-C. Ho, B. Moyers, J. Xu and three anonymous reviewers for comments. Research in the Zhang Lab on the topic reviewed here has been supported by the US National Institutes of Health.

Competing interests statement
The authors declare no competing interests.

SUPPLEMENTARY INFORMATION
See online article: S1 (box) | S2 (box)
ALL LINKS ARE ACTIVE IN THE ONLINE PDF