Published in final edited form as:

Nat Rev Genet. 2015 July ; 16(7): 409–420. doi:10.1038/nrg3950.

Determinants of the rate of protein sequence evolution

Jianzhi Zhang¹ and Jian-Rong Yang¹

¹Department of Ecology and Evolutionary Biology, University of Michigan, 830 North University Avenue, Ann Arbor, MI 48109, USA

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 constitutes functional constraint has remained unclear. The increasing availability of genomic data has allowed for 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 have identified multiple mechanisms behind these observations and demonstrated a prominent role that selection against errors in molecular and cellular processes plays in protein evolution.

The determination of the amino acid sequences of several homologous proteins in the late 1950s and early 1960s were quickly followed by studies estimating 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 a half-century, critical to reconstructing evolutionary history and mechanisms^{4, 5}. Early studies found that different proteins from the same species can evolve at vastly different rates². According to the well accepted explanation by the neutral theory⁶, the rate of protein sequence evolution (k) equals the rate of mutation (μ) multiplied by the proportion (p) of mutations that are neutral, because 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 p . While the role of μ in determining k has been clearly demonstrated⁷, 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 last 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, plays only a minor role⁸, whereas the protein expression level is found to be a major determinant⁹. 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 that natural selection against errors in molecular and cellular processes plays in protein evolution.

We review here the main discoveries made in this journey to characterize the rate of protein evolution. We detail several primary hypotheses and models proposed to explain protein evolution rate and mechanism, as well as relationships between these partially overlapping models. We synthesize the new mechanistic understanding of protein evolution made possible by recent studies based on analysis of large genomic datasets, and offer our views on the significant biological and medical implications of the progress made in this area and major unsolved questions. We focus on evolutionary rate variation among proteins rather than that among sites within a protein¹⁸. We will not discuss the rate variation of a given protein among different species⁷.

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 proposed the molecular clock hypothesis², based on findings of an approximately constant rate of evolution for a given protein across different evolutionary lineages. This discovery opened the door to molecular dating of evolutionary events that could not or did not leave an adequate fossil record and now plays as important a role as paleontology in providing a temporal scale of biological evolution¹⁹. Second, by calculating the evolutionary rates of three proteins, Kimura noticed that the molecular evolutionary rate is too high to have been driven by positive Darwinian selection³, which, in conjunction with other observations in molecular biology²⁰ and population genetic theories, led to the development of the neutral theory⁶, the only paradigm-changing conceptual revolution in evolutionary biology since the maturation of the neo-Darwinism in the 1950s²¹. 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 evolves²². This wide belief is probably attributable to an influential article by Kimura and Ohta over forty years ago that summarized five principles governing molecular evolution²³. 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 over 20 proteins had been estimated. The highest evolutionary rate for a protein was observed in fibrinopeptides and was >1000 times greater than the lowest rate, 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 play critical roles in gene regulation. Nonetheless, they also explained that fibrinopeptides evolve

rapidly because virtually 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 histone, most amino acids at many sites are probably unacceptable because they would affect the histone function²³. Interestingly, this latter explanation was actually referring to a higher functional constraint on histone than fibrinopeptides, not higher functional importance for the former than the latter. These authors were apparently using 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 due 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 1974 paper. She did not answer immediately, but replied 10 days later in an email that it was functional constraint.

Wilson and colleagues²⁴ 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 definition²⁴, 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 if the protein evolutionary rate decreases with the rise of functional importance. However, this test was not feasible until the end of the 20th century when gene deletion became a routine experiment in several model organisms. The first test with large data was conducted by Hurst and Smith²⁵. They classified 175 mouse genes into two groups: essential and nonessential. Essential genes were defined as those that resulted in lethality or infertility when deleted from the mouse genome, whereas deletion of nonessential genes did not. They also measured the nonsynonymous to synonymous substitution rate ratio d_N/d_S (see 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, these authors observed no significant difference in d_N/d_S between essential and nonessential genes and concluded that the functional importance of a protein does not impact its evolutionary rate²⁵.

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 approximately 500 of these gene-deletion *S. cerevisiae* strains, relative to the wild-type strain²⁶. Hirsh and Fraser found, among nonessential genes, 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 gene²⁷.

In the late 1990s, studies were reporting the first genome-wide measurements of gene expression levels using microarray technology²⁸. Based on analysis of these gene expression

data, Pal et al. reported a strong negative correlation between the expression level of a gene and the evolutionary rate of its protein sequence⁹. They subsequently demonstrated that the correlation reported by Hirsh and Fraser disappears after the control of gene expression level²⁹, 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 relation between functional importance and evolutionary rate.

Zhang and He revisited this hypothesis after the availability of additional genomic data from yeast species³⁰. They found a weak but significant negative correlation between gene importance defined by the fitness reduction upon gene deletion and the protein evolutionary rate, with or without controlling gene expression level. Nonetheless, the partial rank correlation, which is the correlation in rank between two variables after the control of confounding factors, indicates that only ~1% of the variance in protein evolutionary rate is explainable 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 study²⁵ with expanded mouse data, reporting a significant negative correlation between gene essentiality and protein evolutionary rate, with or without the control of gene expression level, although the correlation is again weak. In bacteria, protein functional importance significantly impacts protein evolutionary rate³³, but the influence of expression level is much greater³⁴. In summary, experimental studies in both prokaryotes and eukaryotes demonstrated that the functional importance of a protein has only a weak impact on its evolutionary rate.

Why is the correlation between protein functional importance and evolutionary rate so weak? Wang and Zhang⁸ 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, while 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 that in laboratory that is predicted to impact the evolutionary rate. Wang and Zhang⁸ 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 418 laboratory conditions or predicted (for metabolic enzymes) in any of 10,000 simulated nutritional conditions. Even combinations of the above conditions did not enhance the correlation much. Furthermore, they found no significant evolutionary rate difference between enzymes that are essential under any nutritional condition and those that are nonessential 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 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 importance³⁵ and conclude that the prediction is useful³⁶. Wang and Zhang⁸ noted that if they randomly pick two yeast proteins, there is 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 between these two properties. However, when they ranked all proteins by evolutionary rate and compared two proteins that are separated in rank by over 95% of all proteins, the probability that the more slowly evolving one is more 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 for why the rate-importance correlation has been successfully used in predicting the functionality of noncoding 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 human, mouse, and rat³⁶, and completely unconstrained sequences.

Gene expression level

Gene expression level is a major rate determinant

As mentioned above, Pal et al. first reported the unexpected finding in yeast that the evolutionary rate of a protein is strongly negatively correlated with its microarray-based mRNA concentration⁹. This negative correlation is often referred to as the E-R anticorrelation, where E stands for gene expression level and R stands for evolutionary rate. The E-R anticorrelation exists in both prokaryotes and eukaryotes^{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 use data collected from mid-log phase of growth under rich media, which presumably reflect average concentrations across cell cycle stages. In multicellular organisms, mRNA concentration data used are typically from the whole organism or averaged from several examined tissues. While the E-R anticorrelation tends to be present regardless of the tissue where gene expressions are measured, the magnitude of the anticorrelation does vary among tissues¹⁴.

Because of the strong correlation between mRNA and protein concentrations^{38, 39}, the negative correlation between protein concentration and evolutionary rate is also strong⁴⁰. Drummond et al. showed that the E-R anticorrelation is weaker when E stands for protein concentration than when it stands for mRNA concentration¹⁰. However, it is unclear whether this disparity is genuine or simply reflects different qualities of proteomic and transcriptomic data. Because proteomic data exist in much fewer species than do 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} (Supplementary information S1 (box)), and the impact of the expression level of a gene on its evolution transcends the sequence level (Box 3).

The protein misfolding avoidance hypothesis

What is the mechanism underlying the E-R anticorrelation? This anticorrelation cannot be a byproduct of the covariations of protein functional importance with both E and R, because the E-R anticorrelation remains strong after the control of protein functional importance^{30, 31}. Drummond et al. proposed the translational robustness hypothesis to explain the origin of the E-R anticorrelation¹⁰. The key assumption of the hypothesis is that protein misfolding is cytotoxic and reduces fitness⁴³. Because protein translation is not error-free⁴⁴ 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 selection will result in more stable protein structures, lower translational errors, and lower evolutionary rates for more highly expressed proteins¹⁴. 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 gene⁴⁵. In support of the translational robustness hypothesis, Drummond and Wilke found that the favorable use of preferred codons at conserved sites is stronger for the 10% most highly expressed genes than for average genes¹⁴.

Although Drummond et al.'s model encompasses misfolding of correctly translated and mistranslated proteins¹⁴, their studies focused on mistranslation-induced misfolding^{10, 14}. Yang et al. estimated that, depending on the folding stability of a protein, between 5 and 20% of misfolded protein molecules are correctly translated¹³. Their simulation confirmed that the E-R anticorrelation can arise from selection against both translational error-induced and error-free misfolding, prompting the renaming of the translational robustness hypothesis to the protein misfolding avoidance hypothesis¹³ (Fig. 2). The misfolding avoidance hypothesis makes three predictions, each of which have been empirically supported¹³. First, highly expressed proteins were found to have higher folding stabilities than lowly expressed ones¹³. Further, highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins, which are known to be particularly stable⁴⁶. Second, amino acids and codons that increase protein stability were reported to be more prevalent in highly expressed genes than lowly expressed ones¹³. 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 protein¹³.

The protein misinteraction avoidance hypothesis

It is well known that amino acid residues located inside a protein structure, or core residues, play more central roles than surface residues in protein folding stability⁴⁷. However, surface

residues show an E-R anticorrelation as strong as core residues, suggesting that selective pressures other than misfolding avoidance might also work, especially on surface residues¹¹. Yang et al. 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 cores¹¹. Considering the special importance of surface residues in protein-protein interaction, these authors proposed the protein misinteraction avoidance hypothesis¹¹. The 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 3D lattice protein model, Yang et al. confirmed that selection against deleterious misinteraction results in an E-R anticorrelation¹¹. The misinteraction avoidance hypothesis predicts that, compared with lowly expressed proteins, highly expressed proteins disfavor 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 yeast¹¹, *Escherichia coli*⁵¹, and human⁵¹. Yang et al. further predicted that selection against misinteraction should result in translational robustness manifested by reduced mistranslation and reduced misinteraction upon mistranslation¹¹, 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 protein surfaces¹¹. Nevertheless, even together, the two hypotheses seem insufficient to fully explain the anticorrelation, because each of them explains only a moderate fraction of the anticorrelation¹¹.

The mRNA folding requirement hypothesis

Park et al. proposed the mRNA folding requirement hypothesis¹² to explain the E-R anticorrelation. It had been reported that the mRNAs of highly expressed genes have stronger folding (i.e., with more negative free energies or higher fractions of paired bases) than those of lowly expressed ones⁵². Park et al. showed that this disparity is not a byproduct of nucleotide, codon, or amino acid compositional differences among genes of different expression levels, but results from intensified selection for strong mRNA folding in genes of high expressions¹². If high-concentration mRNAs have been selected for strong folding, 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, the lower the substitution rate, creating an E-R anticorrelation (Fig. 2). In support of this hypothesis, Park et al. detected a strong, negative correlation between mRNA folding strength and protein evolutionary rate, both before and after the control of gene expression level¹².

But, why would more highly expressed genes be under selection for stronger mRNAs 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 tradeoff¹⁵. 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 modeled the cost and benefit of fast elongation and found that the optimal ribosomal elongation speed decreases as the expression level of the protein increases¹⁵. 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 Cherry¹⁷ independently proposed the expression cost hypothesis to explain the E-R anticorrelation (Fig. 3). The hypothesis assumes that (i) protein synthesis has associated cost (C) and benefit (B) that are both increasing functions of protein abundance, and (ii) 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 protein abundance, the extra benefit should equal the extra cost, or $B'(\varepsilon) = C'(\varepsilon)$, where the prime symbol stands for 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 optimal protein abundance 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 where 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 functional importance.

The validity of the elegant expression cost hypothesis in explaining the E-R anticorrelation, however, has not been extensively investigated empirically. One piece of evidence used to support the hypothesis is that deleting one allele of a highly expressed gene from a diploid yeast tends to cause more harm than deleting one allele of a lowly expressed gene¹⁶. But it is unclear whether this phenomenon results from the expression cost hypothesis or simply a byproduct 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 $q \ll 1$.

What constituents are included in the cost of protein expression is another critical question. It certainly should include the synthetic material and energy costs of transcription and translation, which is 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 (i) slower evolution of more highly expressed proteins upon the control of protein length and (ii) slower evolution of longer proteins upon the control of 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 costs⁵³. 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 scrutiny.

Correlates of protein evolutionary rate

In addition to the factors discussed, 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 rates⁵⁴. Wolf et al. 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 proteins⁵⁴. This suggests the importance of domain-specific features in determining the protein evolutionary rate, but what these features are is unclear. 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 appear to show moderate impacts. Regardless, the mechanisms behind their direct or indirect impacts are often unknown. For instance, the number of types of microRNAs targeting a mammalian gene is the best predictor of the protein evolutionary rate of the gene⁵⁵, and its impact goes beyond those of gene expression level⁵⁵ and 3' untranslated region length⁵⁶. One suggested mechanism is that the microRNA type number reflects the target's pleiotropic level⁵⁵, which is known to constrain protein evolution⁵⁷. But 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 are both elusive. Furthermore, if pleiotropy is a primary rate determinant, it is puzzling why the number of protein interaction partners that a protein has, 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 factors discussed thus far impact the intensity of purifying selection, which prevents the fixation of (deleterious) mutations. In theory, the rate of protein evolution is also impacted 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 that are subject to recurrent positive selection typically related to host-pathogen interactions⁶⁰ or intersexual interactions⁶¹. 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 "relating to toxicity (or negative function)". In other words, mutations can be unacceptable due to the disruption of a physiological function or 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 does not reflect much 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 fruit fly genes, and the average evolutionary rate of the translated post-stop regions is slightly lower than that of neutral sequences⁶². This observation is not a proof that the post-stop regions have physiological functions, because the observation could be due to toxicity-avoidance constraints. Third, studies over the last decade have revealed many stochastic errors and noises in cellular and molecular processes such as gene expression⁶³ and pre-mRNA splicing⁶⁴. This is hardly surprising, because these processes require biochemical reactions, which are stochastic in nature. The biological significance of these errors and noises is just starting to be unraveled^{65–69}, and the discovery of their dominant roles in shaping protein evolution points to a potential of their involvements in all corners of cellular and molecular biology as well as evolution.

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 a protein is known to cause a number of diseases⁷⁰.

Computational analysis has suggested that up to 80% of disease-causing missense mutations reduce protein structural stability, which would increase the misfolding probability⁷¹. 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 tumor suppressor gene *tp53*⁷². Furthermore, over-expression of a promiscuous protein whose normal expression is low could induce disease-causing protein misinteraction, as demonstrated in cancers⁴⁹. Second, predicting 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 as well as 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 aging^{75, 76}, cancer⁷⁷, and other diseases⁷⁸ 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. While these recent studies have uncovered unsuspected forces in protein evolution, a number of questions remain for future studies. First, although theoretically attractive, the expression cost hypothesis of the E-R anticorrelation 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. Furthermore, 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, translational 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 role in determining the rate of protein sequence evolution (see Box 1), but the details of this distribution have not been worked out 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 enhancing its cytotoxicity. Fifth, because cellular errors may by chance create new protein variants, it would be interesting to study if these errors play important roles in the origin of new protein functions and adaptation⁸¹. 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 noncoding 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 unknown⁸⁵. Finally, the

present 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 promises to offer further insights into the mechanisms of evolution and disease.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Xiaoshu Chen, Wei-Chin Ho, Bryan Moyers, Jinrui Xu, and three anonymous reviewers for constructive comments. Researches in the Zhang lab on the topic reviewed here have been supported by the U.S. National Institutes of Health.

Glossary

Designability	The number of protein sequences that adopt a protein structure
Dispensability	The degree to which an organism can survive and reproduce when a given gene is removed
Effective population size (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, sex ratio, and other factors
Functional constraint	The extent to which random mutations are purged by natural selection due to their deleterious effects on the protein function
Functional importance	The fitness advantage to an organism conferred by the function of a protein
Mistranslation	Incorporation of incorrect amino acids in a nascent protein during synthesis, which may be caused by incorrect charging of tRNAs by aminoacyl tRNA synthetases or incorrect acceptance of tRNAs by ribosomes
Molecular clock	The hypothesis that the same protein evolves with an approximately constant rate over time and across different organisms
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
RNA folding strength	A measure of the reduction in free energy of a folded RNA molecule, compared to its unfolded form
Orthologous genes	Genes from different species that originated by vertical descent from a single gene of the last common ancestor of these species

Pleiotropy	The phenomenon of one gene or one mutation affecting multiple traits
Preferred codons	Codons that are used more frequently than their synonymous codons in a genome sequence
Protein conformational diversity	The degree of structural variations of various native states of a protein
Protein misfolding	The process by which a protein structure assumes its nonnative shape or conformation, which not only diminishes the protein's physiological function but may also create cytotoxicity
Protein misinteraction	A nonnative interaction between protein molecules, which not only reduces the concentrations of freely available protein molecules but may also be toxic

Biographies

Jianzhi Zhang is the Marshall W. Nirenberg Collegiate Professor of Ecology and Evolutionary Biology at University of Michigan. He is interested in molecular and cellular properties that set fundamental constraints of life yet permit and channel evolution at higher levels. He received his undergraduate degree from Fudan University and his doctorate from Pennsylvania State University.

Jian-Rong Yang is a postdoctoral researcher in Prof. Jianzhi Zhang's group at University of Michigan. He is interested in general principles of molecular and genomic evolution, especially in relation to the origin of systemic characteristics of life. He received B.S. and Ph.D. degrees from Sun Yat-Sen University in China.

References

1. Zuckerkandl, E.; Pauling, L. Horizons in Biochemistry. Kasha, M.; Pullman, B., editors. New York: Academic Press; 1962. p. 189-225.
2. Zuckerkandl, E.; Pauling, L. Evolving Genes and Proteins. Bryson, V.; Vogel, HJ., editors. New York: Academic Press; 1965. p. 97-166.
3. Kimura M. Evolutionary rate at the molecular level. *Nature*. 1968; 217:624–626. [PubMed: 5637732]
4. Kumar S. Molecular clocks: four decades of evolution. *Nat Rev Genet*. 2005; 6:654–662. [PubMed: 16136655]
5. Takahata N. Molecular clock: an anti-neo-Darwinian legacy. *Genetics*. 2007; 176:1–6. [PubMed: 17513888]
6. Kimura, M. The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press; 1983.
7. Li, W. Molecular Evolution. Sinauer: Sunderland, Mass; 1997.
8. Wang Z, Zhang J. Why is the correlation between gene importance and gene evolutionary rate so weak? *PLoS Genet*. 2009; 5:e1000329. [PubMed: 19132081]
9. Pal C, Papp B, Hurst LD. Highly expressed genes in yeast evolve slowly. *Genetics*. 2001; 158:927–931. [PubMed: 11430355] First report of the E-R anticorrelation.

10. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*. 2005; 102:14338–14343. [PubMed: 16176987] Proposes the translational robustness hypothesis of the E-R anticorrelation.
11. Yang JR, Liao BY, Zhuang SM, Zhang J. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc Natl Acad Sci U S A*. 2012; 109:E831–E840. [PubMed: 22416125] Proposes the protein misinteraction hypothesis of the E-R anticorrelation.
12. Park C, Chen X, Yang JR, Zhang J. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*. 2013; 110:E678–E686. [PubMed: 23382244] Proposes the mRNA folding requirement hypothesis of the E-R anticorrelation.
13. Yang JR, Zhuang SM, Zhang J. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol Syst Biol*. 2010; 6:421. [PubMed: 20959819]
14. Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*. 2008; 134:341–352. [PubMed: 18662548]
15. Yang JR, Chen X, Zhang J. Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS Biol*. 2014; 12:e1001910. [PubMed: 25051069] Explains the underlying cause of the mRNA folding requirement that partially accounts for the E-R anticorrelation.
16. Gout JF, Kahn D, Duret L. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet*. 2010; 6:e1000944. [PubMed: 20485561] Proposes the expression cost hypothesis of the E-R anticorrelation.
17. Cherry JL. Expression level, evolutionary rate, and the cost of expression. *Genome Biol Evol*. 2010; 2:757–769. [PubMed: 20884723] Independently proposes 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*. 1996; 11:367–372. [PubMed: 21237881]
19. Hedges SB, Dudley J, Kumar S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*. 2006; 22:2971–3972. [PubMed: 17021158]
20. King JL, Jukes TH. Non-Darwinian evolution. *Science*. 1969; 164:788–798. [PubMed: 5767777]
21. Zhang, J. *Evolution Since Darwin: The First 150 Years*. Bell, MA.; Futuyma, DJ.; Eanes, WF.; Levinton, JS., editors. Sinauer: Sunderland, Mass; 2010. p. 87-118.
22. Karp, G. *Cell and Molecular Biology*. Hoboken, New Jersey: John Wiley & Sons, Inc; 2008.
23. Kimura M, Ohta T. On some principles governing molecular evolution. *Proc Natl Acad Sci U S A*. 1974; 71:2848–2852. [PubMed: 4527913] Proposes the role of protein functional importance and functional constraint in determining the rate of protein sequence evolution.
24. Wilson AC, Carlson SS, White TJ. Biochemical evolution. *Annu Rev Biochem*. 1977; 46:573–639. [PubMed: 409339]
25. Hurst LD, Smith NG. Do essential genes evolve slowly? *Curr Biol*. 1999; 9:747–750. [PubMed: 10421576] First test of the relationship between protein functional importance and evolutionary rate based on a relatively large genomic dataset.
26. Winzeler EA, et al. Functional characterization of the *Scerevisiae* genome by gene deletion and parallel analysis. *Science*. 1999; 285:901–906. [PubMed: 10436161]
27. Hirsh AE, Fraser HB. Protein dispensability and rate of evolution. *Nature*. 2001; 411:1046–1049. [PubMed: 11429604]
28. Holstege FC, et al. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*. 1998; 95:717–728. [PubMed: 9845373]
29. Pal C, Papp B, Hurst LD. Genomic function: Rate of evolution and gene dispensability. *Nature*. 2003; 421:496–497. discussion 497–8. [PubMed: 12556881]
30. Zhang J, He X. Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol Biol Evol*. 2005; 22:1147–1155. [PubMed: 15689524]
31. Wall DP, et al. Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci U S A*. 2005; 102:5483–5488. [PubMed: 15800036]

32. Liao BY, Scott NM, Zhang J. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Mol Biol Evol.* 2006; 23:2072–2080. [PubMed: 16887903]
33. Jordan IK, Rogozin IB, Wolf YI, Koonin EV. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* 2002; 12:962–968. [PubMed: 12045149]
34. Rocha EP, Danchin A. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol.* 2004; 21:108–116. [PubMed: 14595100]
35. Lindblad-Toh K, et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature.* 2011; 478:476–482. [PubMed: 21993624]
36. Pennacchio LA, et al. In vivo enhancer analysis of human conserved non-coding sequences. *Nature.* 2006; 444:499–502. [PubMed: 17086198]
37. Krylov DM, Wolf YI, Rogozin IB, Koonin EV. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* 2003; 13:2229–2235. [PubMed: 14525925]
38. Greenbaum D, Colangelo C, Williams K, Gerstein M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.* 2003; 4:117. [PubMed: 12952525]
39. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet.* 2012; 13:227–2232. [PubMed: 22411467]
40. Drummond DA, Raval A, Wilke CO. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol.* 2006; 23:327–337. [PubMed: 16237209]
41. Shen Y, et al. Testing hypotheses on the rate of molecular evolution in relation to gene expression using microRNAs. *Proc Natl Acad Sci U S A.* 2011; 108:15942–15947. [PubMed: 21911382]
42. Managadze D, Rogozin IB, Chernikova D, Shabalina SA, Koonin EV. Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol Evol.* 2011; 3:1390–1404. [PubMed: 22071789]
43. Geiler-Samerotte KA, et al. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc Natl Acad Sci U S A.* 2011; 108:680–685. [PubMed: 21187411]
44. Drummond DA, Wilke CO. The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet.* 2009; 10:715–724. [PubMed: 19763154]
45. Akashi H. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics.* 1994; 136:927–935. [PubMed: 8005445]
46. Cherry JL. Highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins. *Mol Biol Evol.* 2010; 27:735–741. [PubMed: 19910385]
47. Chakravarty S, Varadarajan R. Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure.* 1999; 7:723–732. [PubMed: 10425675]
48. Stambolsky P, et al. Modulation of the vitamin D3 response by cancer-associated mutant p53. *Cancer Cell.* 2010; 17:273–285. [PubMed: 20227041]
49. Vavouri T, Semple JI, Garcia-Verdugo R, Lehner B. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell.* 2009; 138:198–208. [PubMed: 19596244]
50. Zhang J, Maslov S, Shakhnovich EI. Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size. *Mol Syst Biol.* 2008; 4:210. [PubMed: 18682700]
51. Levy ED, De S, Teichmann SA. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proc Natl Acad Sci U S A.* 2012; 109:20461–20466. [PubMed: 23184996]
52. Zur H, Tuller T. Strong association between mRNA folding strength and protein abundance in *S cerevisiae*. *EMBO Rep.* 2012; 13:272–277. [PubMed: 22249164]
53. Akashi H, Gojobori T. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc Natl Acad Sci U S A.* 2002; 99:3695–3700. [PubMed: 11904428]

54. Wolf MY, Wolf YI, Koonin EV. Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution. *Biol Direct*. 2008; 3:40. [PubMed: 18840284]
55. Chen SC, Chuang TJ, Li WH. The relationships among microRNA regulation, intrinsically disordered regions, and other indicators of protein evolutionary rate. *Mol Biol Evol*. 2011; 28:2513–2520. [PubMed: 21398349]
56. Cheng C, Bhardwaj N, Gerstein M. The relationship between the evolution of microRNA targets and the length of their UTRs. *BMC Genomics*. 2009; 10:431. [PubMed: 19751524]
57. He X, Zhang J. Toward a molecular understanding of pleiotropy. *Genetics*. 2006; 173:1885–1891. [PubMed: 16702416]
58. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. *Science*. 2002; 296:750–752. [PubMed: 11976460]
59. Bloom JD, 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*. 2003; 3:21. [PubMed: 14525624]
60. Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*. 1988; 335:167–170. [PubMed: 3412472]
61. Lee YH, Ota T, Vacquier VD. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Mol Biol Evol*. 1995; 12:231–238. [PubMed: 7700151]
62. Dunn JG, Foo CK, Belletier NG, Gavis ER, Weissman JS. Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster*. *Elife*. 2013; 2:e01179. [PubMed: 24302569]
63. Raj Oudenaarden A, van A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*. 2008; 135:216–226. [PubMed: 18957198]
64. Marinov GK, et al. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res*. 2014; 24:496–510. [PubMed: 24299736]
65. Zhang Z, Qian W, Zhang J. Positive selection for elevated gene expression noise in yeast. *Mol Syst Biol*. 2009; 5:299. [PubMed: 19690568]
66. Wang Z, Zhang J. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proc Natl Acad Sci U S A*. 2011; 108:E67–E76. [PubMed: 21464323]
67. Warnecke T, Hurst LD. Error prevention and mitigation as forces in the evolution of genes and genomes. *Nat Rev Genet*. 2011; 12:875–881. [PubMed: 22094950]
68. Eldar A, Elowitz MB. Functional roles for noise in genetic circuits. *Nature*. 2010; 467:167–173. [PubMed: 20829787]
69. Xu G, Zhang J. Human coding RNA editing is generally nonadaptive. *Proc Natl Acad Sci U S A*. 2014; 111:3769–3774. [PubMed: 24567376]
70. Gregersen N, Bross P, Vang S, Christensen JH. Protein misfolding and human disease. *Annu Rev Genomics Hum Genet*. 2006; 7:103–124. [PubMed: 16722804]
71. Wang Z, Moult J. SNPs, protein structure, and disease. *Hum Mutat*. 2001; 17:263–270. [PubMed: 11295823]
72. Oren M, Rotter V. Mutant p53 gain-of-function in cancer. *Cold Spring Harb Perspect Biol*. 2010; 2:a001107. [PubMed: 20182618]
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*. 2014; 10:e1004237. [PubMed: 24651380]
74. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet*. 2011; 12:628–640. [PubMed: 21850043]
75. Orgel LE. The maintenance of the accuracy of protein synthesis and its relevance to ageing. *Proc Natl Acad Sci U S A*. 1963; 49:517–521. [PubMed: 13940312]
76. Silva RM, et al. The yeast PNC1 longevity gene is up-regulated by mRNA mistranslation. *PLoS One*. 2009; 4:e5212. [PubMed: 19381334]
77. Pandolfi PP. Aberrant mRNA translation in cancer pathogenesis: an old concept revisited comes finally of age. *Oncogene*. 2004; 23:3134–3137. [PubMed: 15094762]

78. Frank SA. Somatic mosaicism and disease. *Curr Biol*. 2014; 24:R577–R581. [PubMed: 24937287]
79. Stiffler MA, Hekstra DR, Ranganathan R. Evolvability as a function of purifying selection in TEM-1 beta-lactamase. *Cell*. 2015; 160:882–892. [PubMed: 25723163]
80. Podgornaia AI, Laub MT. Pervasive degeneracy and epistasis in a protein-protein interface. *Science*. 2015; 347:673–677. [PubMed: 25657251]
81. Whitehead DJ, Wilke CO, Vernazobres D, Bornberg-Bauer E. The look-ahead effect of phenotypic mutations. *Biol Direct*. 2008; 3:18. [PubMed: 18479505]
82. Pal C, Papp B, Lercher MJ. An integrated view of protein evolution. *Nat Rev Genet*. 2006; 7:337–348. [PubMed: 16619049]
83. Wolf YI, Carmel L, Koonin EV. Unifying measures of gene function and evolution. *Proc Biol Sci*. 2006; 273:1507–1515. [PubMed: 16777745]
84. Xia Y, Franzosa EA, Gerstein MB. Integrated assessment of genomic correlates of protein evolutionary rate. *PLoS Comput Biol*. 2009; 5:e1000413. [PubMed: 19521505]
85. Du X, Lipman DJ, Cherry JL. Why does a protein's evolutionary rate vary over time? *Genome Biol Evol*. 2013; 5:494–503. [PubMed: 23436005]
86. Franzosa EA, Xia Y. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol*. 2009; 26:2387–2395. [PubMed: 19597162]
87. Yeh SW, et al. Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure. *Mol Biol Evol*. 2014; 31:135–129. [PubMed: 24109601]
88. Zhang J, Gu X. Correlation between the substitution rate and rate variation among sites in protein evolution. *Genetics*. 1998; 149:1615–1625. [PubMed: 9649548]
89. Chen FC, Liao BY, Pan CL, Lin HY, Chang AY. Assessing determinants of exonic evolutionary rates in mammals. *Mol Biol Evol*. 2012; 29:3121–3129. [PubMed: 22504521]
90. Nei, M.; Kumar, S. *Molecular Evolution and Phylogenetics*. New York: Oxford University Press; 2000.
91. Kim N, Jinks-Robertson S. Transcription as a source of genome instability. *Nat Rev Genet*. 2012; 13:204–214. [PubMed: 22330764]
92. Hanawalt PC, Spivak G. Transcription-coupled DNA repair: two decades of progress and surprises. *Nat Rev Mol Cell Biol*. 2008; 9:958–970. [PubMed: 19023283]
93. Park C, Qian W, Zhang J. Genomic evidence for elevated mutation rates in highly expressed genes. *EMBO Rep*. 2012; 13:1123–1129. [PubMed: 23146897]
94. Chen X, Zhang J. No gene-specific optimization of mutation rate in *Escherichia coli*. *Mol Biol Evol*. 2013; 30:1559–1562. [PubMed: 23533222]
95. Chen X, Zhang J. Yeast mutation accumulation experiment supports elevated mutation rates at highly transcribed sites. *Proc Natl Acad Sci U S A*. 2014; 111:E4062. [PubMed: 25217566]
96. Lind PA, Andersson DI. Whole-genome mutational biases in bacteria. *Proc Natl Acad Sci U S A*. 2008; 105:17878–17883. [PubMed: 19001264]
97. Gottipati P, Helleday T. Transcription-associated recombination in eukaryotes: link between transcription, replication and recombination. *Mutagenesis*. 2009; 24:203–210. [PubMed: 19139058]
98. Liao BY, Zhang J. Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Mol Biol Evol*. 2006; 23:1119–1128. [PubMed: 16520335]
99. Park C, Zhang J. High expression hampers horizontal gene transfer. *Genome Biol Evol*. 2012; 4:523–532. [PubMed: 22436996]
100. Zhang L, Li WH. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol Biol Evol*. 2004; 21:236–239. [PubMed: 14595094]
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*. 2013; 9:e1003476. [PubMed: 23637639]

102. Chuang TJ, Chiang TW. Impacts of pretranscriptional DNA methylation, transcriptional transcription factor, and posttranscriptional microRNA regulations on protein evolutionary rate. *Genome Biol Evol.* 2014; 6:1530–1541. [PubMed: 24923326]
103. Taipale M, et al. Quantitative analysis of HSP90-client interactions reveals principles of substrate recognition. *Cell.* 2012; 150:987–1001. [PubMed: 22939624]
104. Bogumil D, Dagan T. Chaperonin-dependent accelerated substitution rates in prokaryotes. *Genome Biol Evol.* 2010; 2:602–608. [PubMed: 20660111]
105. Liao BY, Weng MP, Zhang J. Impact of extracellularly on the evolutionary rate of mammalian proteins. *Genome Biol Evol.* 2010; 2:39–43. [PubMed: 20333223]
106. Sharp PM, Shields DC, Wolfe KH, Li WH. Chromosomal location and evolutionary rate variation in enterobacterial genes. *Science.* 1989; 246:808–810. [PubMed: 2683084]
107. Flynn KM, Vohr SH, Hatcher PJ, Cooper VS. Evolutionary rates and gene dispensability associate with replication timing in the archaeon *Sulfolobus islandicus*. *Genome Biol Evol.* 2010; 2:859–869. [PubMed: 20978102]
108. Hahn MW, Kern AD. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol.* 2005; 22:803–806. [PubMed: 15616139]
109. Kim PM, Lu LJ, Xia Y, Gerstein MB. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science.* 2006; 314:1938–1941. [PubMed: 17185604]
110. Vitkup D, Kharchenko P, Wagner A. Influence of metabolic network structure and function on enzyme evolution. *Genome Biol.* 2006; 7:R39. [PubMed: 16684370]
111. Jovelín R, Phillips PC. Evolutionary rates and centrality in the yeast gene regulatory network. *Genome Biol.* 2009; 10:R35. [PubMed: 19358738]
112. Kryuchkova N, Robinson-Rechavi M. Determinants of protein evolutionary rates in light of ENCODE functional genomics. *BMC Bioinformatics.* 2014; 15:A9.
113. Bloom JD, Drummond DA, Arnold FH, Wilke CO. Structural determinants of the rate of protein evolution in yeast. *Mol Biol Evol.* 2006; 23:1751–1761. [PubMed: 16782762]
114. Brown CJ, et al. Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol.* 2002; 55:104–110. [PubMed: 12165847]
115. Javier Zea D, Miguel Monzon A, Fornasari MS, Marino-Buslje C, Parisi G. Protein conformational diversity correlates with evolutionary rate. *Mol Biol Evol.* 2013; 30:1500–1503. [PubMed: 23564939]

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 where the proteins reside (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 in wide use⁹⁰. 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 among-protein 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 site⁹⁰. Because d_S is largely determined by mutation rate only, whereas d_N is determined jointly by mutation rate and selection, d_N/d_S is determined by selection only.

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

Qualitatively speaking, 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. Let us first 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 organism's population size, and N_e be the effective population size. The protein substitution rate in diploids is $k = (1 - \alpha)\mu + \alpha(2N\mu)f = \mu[1 -$

$\alpha(1 - 2Nf)]$, where $f = \frac{1 - e^{-\beta N_e/N}}{1 - e^{-2\beta N_e}}$ is the fixation probability of a new null mutation⁶. It can be shown⁸ 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 either null or have no functional effect, a strong negative correlation between functional importance and evolutionary rate is not expected⁸. A hypothetical scenario with $\alpha = 0.8$ is shown in the top row of the figure, where the left panel depicts the cumulative probability distribution of deleterious functional effects of random mutations, the middle panel depicts the theoretical relationship between the functional importance of a gene and the protein evolutionary rate measured by d_N/d_S , and the right panel depicts the same relationship for 1000 genes simulated using functional importance and population size data from the budding yeast with the consideration of estimation errors. Nonetheless, under a different model with the presence of a sizable fraction of deleterious mutations whose functional effects are between $1/N_e$ and $100/N_e$, a substantial correlation between functional importance and protein evolutionary rate becomes possible⁸. A hypothetical scenario is shown in the bottom row of the figure, with 20% of null mutations, 20% neutral mutations, and 60% slightly deleterious mutations whose functional effects follow the beta distribution of $\beta(1, 10^6)$. [The figure is adapted with permission from Ref. ⁸.]

Box 3. Impacts of gene expression level on other aspects of molecular evolution

In addition to impacting 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 germline 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 are opposite, the selective effect is expected to be stronger than what the current E-R anticorrelation reveals⁹³. Second, transcription is known to induce recombination⁹⁷. Third, high gene expression is correlated with a low rate of among-tissue expression-profile evolution in mammals, although the mechanism remains unclear⁹⁸. Fourth, highly expressed genes are more resistant than lowly expressed genes to gene dosage changes¹⁶. 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 (1) energy expenditure in transcription and translation, (2) cytotoxic protein misfolding, (3) reduction in cellular translational efficiency, (4) detrimental protein misinteraction, and/or (5) 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 level⁹⁹.

Online Summary

- Studying the rate of protein sequence evolution led to the foundation of the field of molecular evolution and continues to offer insights on the mechanism of evolution.
- The evolutionary rate of a protein is only weakly influenced by the functional importance of the protein.
- The expression level of a protein is a major determinant of the evolutionary rate of the protein.
- Natural selection against several molecular and cellular errors such as mistranslation, protein misfolding, and protein misinteraction is a primary explanation of why highly expressed proteins evolve slowly.
- The functional constraints on a protein include not only a constraint to maintain its physiological function but also a constraint to avoid toxicity, and both factors influence the evolutionary rate of the protein.

Editorial Summary

The rate and mechanism of protein sequence evolution are fundamental questions within the field of molecular evolution. In this Review, the authors examine theoretical models as well as empirical testing based on recent analysis of large-scale genomic datasets that have offered new insights into the determinants of the rate of protein evolution.

Fig. 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 gray dot represents one gene and the color (red to light blue) represents 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.

Fig. 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 and the types of errors generated in these processes. Red stars indicate translational errors. The lower part of the figure shows expected relationships between the expression level of a protein and 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.

Fig. 3. The expression cost hypothesis of the E-R anticorrelation

Each blue curve represents the benefit of the expression of a gene, whereas the red line shows the cost of expression for each gene. The top half of the figure shows that the benefit and 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). The bottom half of the

figure shows the expected relationships between various gene properties and expression level based on the model.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Correlates of the protein evolutionary rate

Correlates	Faster evolving proteins have	Organisms	References
Gene expression level	Lower expressions	Prokaryotes 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	Expressions in late embryogenesis and adult	Zebrafish	101
Promoter and gene body methylation	Higher promoter methylation but lower gene body methylation	Mammals	102
Chaperone targeting	Stronger chaperone targeting	Prokaryotes and eukaryotes	103, 104
Protein subcellular localization	Higher tendency to be extracellular	Yeast and mammals	105
Codon usage bias	Weaker codon usage bias	Prokaryotes and eukaryotes	14
Distance from the origin of replication	Larger distance	Prokaryotes	106, 107
Pleiotropy	Lower pleiotropy	Eukaryotes	57
Protein interaction network properties	Lower connectivity, closeness, and betweenness	Eukaryotes	58, 108, 109
Metabolic network property	Lower flux and connectivity	Yeast	110
Regulatory network properties	Higher centrality	Yeast	111
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	112, 113
mRNA folding	Weaker mRNA folding	Prokaryotes and eukaryotes	12
GC%	Lower GC%	Mammals	112
Domain structure	Lower density of domains	Animals and plants	54
Protein disordered regions	More disordered regions	Prokaryotes and eukaryotes	114
Protein structural designability	Higher interresidue contact density and higher fraction of buried sites	Yeast	113
Protein conformational diversity	Lower conformational diversity	Mammals	115