FEBS Letters 583 (2009) 1053–1059

FEBS Letters

journal homepage: www.FEBSLetters.org

Minireview

Why proteins evolve at different rates: The functional hypothesis versus the mistranslation-induced protein misfolding hypothesis

Donghyun Park[a], Sun Shim Choi[b,c,*]

[a] Howard Huges Medical Institute, Department of Human Genetics, University of Chicago, Chicago, IL, USA
[b] Department of Molecular and Medical Biotechnology, Kangwon National University, Chunchon 200-701, Republic of Korea
[c] Institute of Bioscience and Biotechnology, Kangwon National University, Chunchon 200-701, Republic of Korea

ARTICLE INFO

Article history:
Received 31 December 2008
Revised 6 February 2009
Accepted 25 February 2009
Available online 28 February 2009

Edited by Takashi Gojobori

Keywords:
Protein evolutionary rate
Functional hypothesis
Mistranslation-induced protein misfolding hypothesis
Translational robustness
Expression abundance

ABSTRACT

Protein evolutionary rates have been presumed to be mostly determined by the density of functionally important amino acids in a protein. They have been shown to correlate with variables intuitively related to functional importance of proteins, such as protein dispensability and protein–protein interactions. Surprisingly, the best correlate of the evolutionary rates has turned out to be not the functional importance of a protein but the expression level of the protein. Drummond and Wilke suggest that the dominant role of expression levels in slowing the rate of protein evolution stems from a selection pressure against mistranslation-induced protein misfolding. We will review current evidence for and against the different hypotheses on determining evolutionary rates.

1. Introduction

The first grand generalization of molecular evolution is that proteins evolve at widely different rates but each particular protein has a characteristic rate that remains relatively constant over long evolutionary spans [1]. In other words, there seems to be a molecular clock that ticks at widely different paces for different protein-coding genes. What determines this characteristic rate is one of the central questions in evolutionary biology. Thirty years ago, Zuckerkandl [2] proposed that a protein's sequence will evolve at a rate primarily determined by the proportion of its sites involved in specific functions (or "functional density"): functional constraints dictate protein evolutionary rates ("functional hypothesis") (Table 1). It was an intuitively plausible explanation although testing this proposal at the time was hardly feasible, given that the functions and structures of proteins are indeed widely different and so are the rates of sequence evolution. Despite wide acceptance of the idea that functional constraints dictate protein evolutionary rates, the measurement of functional density remains problematic because residues may contribute to protein function in unpredictable ways, and arduous sequence-wide saturation mutagenesis and mutant characterization studies are required to ascertain these effects.

In the era of systems biology, various types of genome-scale datasets allow us to elucidate determinants of protein evolutionary rate, which has been actively debated over the past several decades with little empirical data. Comparative analysis of sequence data contributes to demonstrate some general idea for the similarities or differences in determining the protein evolutionary rates among different species. Moreover, a different kind of genome-wide information is becoming increasingly available, which includes gene expression level, protein–protein interactions, regulatory network structure and the effect of gene knockout on the organism's fitness. A large increase in the amount of available genome-scale data in the past few years prompted a basic level of analysis in evolutionary systems biology that involves identification of correlations between diverse genome-wide variables, and many such correlations have been described (Table 2). More often than not, however, the interpretation of these observations remains problematic for at least two reasons. First, although statistically significant thanks to the huge number of data points, the correlations are usually relatively weak. Second, the existence of multiple weak correlations makes it hard to identify the primary or causative variables. Recently, multivariate analyses have been performed to uncover primary correlations [3–6]. One of the interesting conclusions is that

* Corresponding author. Address: Department of Molecular and Medical Biotechnology, Institute of Bioscience and Biotechnology, Kangwon National University, Chunchon 200-701, Republic of Korea. Fax: +82 33 241 6480.
E-mail address: schoi@kangwon.ac.kr (S.S. Choi).

0014-5793/$36.00 © 2009 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2009.02.033

Table 1
Comparison of two hypotheses.

	Functional hypothesis	MIM hypothesis
Fitness cost	Abnormal protein function	Cytotoxicity due to misfolded protein
Selection on synonymous mutations	Invisible	Visible
Correlates	Depend on functional importance	Depend on mRNA level
Abundant proteins	Evolve slowly	Evolve slowly

protein abundance has a far greater effect than other more intuitively appealing factors such as protein dispensability or the number of interaction partners.

In this minireview, we will re-examine the functional hypothesis in conjunction with studies on the correlations of protein evolutionary rates with genome variables, while discussing some pitfalls with the hypothesis. We will then cover new hypotheses which better explain why highly expressed genes evolve slowly and glimpses of biological meaning that are starting to emerge from the new perspective.

2. Correlations of variables with protein evolutionary rates from the perspective of the functional hypothesis

The strength and extent of natural selection on individual amino acids in a protein greatly influences the evolutionary rate of that protein. Strong purifying selection leads to a reduced overall protein evolutionary rate while relaxed selection or strong positive selection leads to a rapid rate of evolution. This has been our paradigm for 30 years: functional constraints dictate protein evolutionary rates ("functional hypothesis"). The functional hypothesis predicts a negative correlation between the severity of a gene knockout effect and its protein evolution rate such that essential genes evolve slowly (Fig. 1). Hurst and Smith were the first to test the hypothesis on a set of mammalian proteins [7]. After excluding fast-evolving immune system genes thought to be subject to positive selection, they concluded that there was no reliable correlation between protein evolutionary rates and the severity of the knockout phenotypes. However, subsequent analyses in yeast and in bacteria reversed this conclusion by demonstrating statistically significant, albeit relatively weak, negative correlations between the strength of a gene's knockout fitness effect and its evolutionary rate [8,9]. The negative results of Hurst and Smith have been attributed primarily to the smaller dataset used in their study. Wall et al. found significant independent correlations between evolutionary rate and protein dispensability (inversely related to the overall importance of a protein approximated by the fitness of the corresponding gene knockout strain under various laboratory conditions) [10]. In addition to yeast, the correlation has been also shown in bacterial species [9] and in *Caenorhabditis elegans* [11].

Protein–protein interaction is another measurement that may approximate functional density of proteins assuming that it constrains interfacial residues [12]. This seemingly provocative link has been reported: the hubs of the network are significantly enriched for essential genes [13]. Fraser and his colleagues have shown that protein evolutionary rate inversely correlates with the number of protein–protein interactions in yeast, *i.e.*, the greater the number of interactions a protein has with other proteins, the slower is its likely evolution [12,14]. The negative correlation of the evolutionary rates with protein–protein interactions was also reported in *C. elegans* and *Drosophila melanogaster* [13,15]. Using curated sets of interacting protein crystal structures, Mintseris and Weng concluded that residues in the interfaces of obligate complexes tend to evolve at a relatively slower rate [16].

Protein evolutionary rates have been also reported to correlate with other genome variables (Table 2) including expression level (or breadth) [10,15,17–23] and a gene's propensity to be lost (computed based on the pattern of presence and absence of genes across multiple genomes) [24]. The functional hypothesis posits that each protein molecule by performing its function contributes a small amount to organism fitness, so mutations that reduce two proteins' functional output (e.g., catalytic rate) equally will have fitness effects weighted by the number of molecules of each protein in the cell, or their abundances, causing the more abundant protein to evolve slower. A gene's propensity to be lost is another intuitive correlate of the dispensability of a gene: if a gene is never lost during evolution that is probably because it is essential for viability. Thus the observed correlation seems to support the idea that functional constraint is a selection pressure causing the variation in the protein evolutionary rates. However, there are some difficulties with the functional hypothesis in explaining whole correlations among different genome variables.

3. What the functional hypothesis cannot explain

First of all, the evolutionary rates show surprisingly weak correlations with several measures of functional importance such as essentiality (functional importance of a protein) and the number of protein–protein interactions. Conflicting results have been published on the validity and the significance of the correlation depending on the datasets and analysis methods [25–30]. Table 2 summarizes references on the correlations that have been reported to exist between protein evolutionary rates and variables. The controversies around the correlations show that it is usually difficult to establish cause–effect associations among many intercorrelated variables, particularly when variables are imprecisely measured and/or can only be measured indirectly through other variables. Specifics of the controversies are not a main focus of this review, but the lesson here is that the complexities and interdependencies of the genome variables must be properly accounted for [30]. An association with the evolutionary rates should only be considered seriously if it holds a significant correlation in different biological systems after controlling confounding effects. To resolve this issue, multivariate analyses have been performed to uncover primary connections [3–6,31,32], demonstrating that the influence of protein–protein interactions and dispensability is decreased when expression level is controlled for [29,33]. A surprising conclusion from these studies is that protein abundance has a far greater effect than other more intuitively appealing variables such as protein dispensability or the number of interaction partners in determining the evolutionary rates [34,35].

Genes with high mRNA expression levels encode slow-evolving proteins, from bacteria [6,36], yeast [17,22], and algae [37] to nematodes [24], plants [20,38], fruit flies [15], mice, and humans [18]. Whereas most variables have little, if any, explanatory power, expression levels account for a significant proportion of the variance in the evolutionary rates of proteins. Expression, measured indirectly using codon usage bias, accounts for ~30% of all variance in protein substitution rates in bacteria [36], ~36% in yeast [17], ~32% in *Chlamydomonas* [37] and ~25% in *Drosophila* [39] (for review, see [35]). Other proxies of expression levels lead to qualitatively similar results. In yeast, the ratio of divergence among paralogues after duplication also depends on expression levels because it correlates with the ratio of mRNA abundances, explaining ~30% of the variance. This is significant evidence for one single variable to be a key element in determining the protein evolutionary rates. It is quite a striking conclusion that, from bacteria to mammals, the best correlate of the evolutionary rates is not functional importance of a protein, but the expression level of the protein. What could justify such a surprising observation?

4. Translational accuracy hypothesis

If expression level has a far greater effect than dispensability and/or essentiality, selection pressures for translational accuracy

Table 2
Known correlates of protein evolutionary rates.

Genome variables	Species	# of total genes analyzed (NE/E)	Datasets for variables	Correlation	Species for divergence calculation (outgoup)	Ref.
Propensity of gene loss	Yeast, worm, human	3140 KOGs		Positive	Seven eukaryotic genomes (*Arabidopsis thaliana*)	[24]
Protein length	Yeast	5865		Positive[a]	*Saccharomyces cerevisiae–Saccharomyces bayanus*	[32]
	Fly	1258		Positive[a]	*Drosophila melanogaster–Drosophila pseudoobscura* (*Anopheles gambiae*)	[15]
	Plant	558		Positive[a]	*Populus tremula-Populus trichocarpa*	[20]
Designability (protein's contact density)	Yeast	5865	Protein databank	Positive[a]	*S. cerevisiae–S. bayanus*	[32]
	Bacteria, yeast, fly, human	777 *Escherichia coli*, 363 *S. cerevisiae*, 795 *melanogaster*, 860 (PDB, GTOP) *Homo sapiens*	*D. protein databank*	Positive[a]	*E. coli–Salmonella typhimurium*, *S. cerevisiae–S. bayanus*, *D. melanogaster–Drosophila yakuba*, *H. sapiens*–Mus musculus	[59]
	Bacteria	~4100 *Bacillus subtilis* ~4300 *E. coli*		Negative	Five species including *E. coli* & *B. subtilis*	[36]
	Yeast	185 gene pairs (duplication study)	Microarray	Negative	*S. cerevisiae–Candida albicans*	[17]
	Yeast	5724	Microarray CAI	Negative	*S. cerevisiae*–9 other yeast species	[19]
	Yeast	3038	Microarray	Negative	Four yeast species	[10]
	Yeast	290	Microarray	Negative	*S. cerevisiae*–4 other yeast species (*Kluyveromyces waltii*)	[22]
Expression level (mRNA)	Fly	1258	Microarray	Negative	*D. pseudoobscura–D. melanogaster* (*A. gambiae*)	[15]
	Human, mouse, rat	7383 Human, 6724 mouse	Microarray	Negative	*H. sapiens–M. musculus*	[60]
	Fly, mouse	60229 *Drosophila* ESTs, [illegible]72 Mouse ESTs	EST, microarray	Negative	Eight species including bacteria, plants, and animals	[18]
	Bacteria, yeast, worm, fly, mouse, human	2229 Bacteria, [illegible]2 yeast, 2[illegible]6 worm, 6649 [illegible], 6167 mouse, 3180 human	Microarray	Negative	*E. coli–Salmonella typhimurium*, *S. cerevisiae–Saccharomyces paradoxus*, *C. elegans–C. briggsae*, *D. melanogaster–D. yakuba*, *M. musculus–Rattus norgegicus*, *H. sapiens*–it Canis familiaris	[21]
	Plant	558	EST	Negative	*Populus tremula–Populus trichocarpa*	[20]
Expression breath	Human, mouse	[illegible] human/rodent, 8[illegible] mouse/[illegible]	EST	Negative	*H. sapiens–M. musculus–Rattus norvegicus*	[23]
	Fly, mouse, human	6649 Fly, [illegible]7 mouse, 3180 human	Microarray	Negative	*E. coli–S. typhimurium*, *S. cerevisiae–S. paradoxus*, *C. elegans–C. briggsae*, *D. melanogaster–D. yakuba*, *M. musculus–R. norgegicus*, *H. sapiens–C. familiaris*	[21]
	Plant	558	EST	Negative[a]	*P. tremula–P. trichocarpa*	[20]
Codon adaptation index (CAI) or frequency of optimal codons (F_{op})	Bacteria	*B. subtilis*-4100, *E. coli*-4300		Negative	Five species including *E. coli* and *B. subtilis*	[36]
	Yeast	3038		Negative	Four yeast species including *S. cerevisiae*	[10]
		548		F_{op} are associated with conserved sites	*C. elegans–H. sapiens*	[39]
	Bacteria, yeast, worm, fly, mouse, human	2786 Bacteria, 4616 yeast, 4173 worm, 7070 fly, 9061 mouse, 5939 human		Negative	*E. coli–S. typhimurium*, *S. cerevisiae - S. paradoxus*, *C. elegans–C. briggsae*, *D. melanogaster–D. yakuba*, *M. musculus–R. norgegicus*, *H. sapiens–C. familiaris*	[21]
	Algae	67	EST	Negative	*Chlamydomonas incerta–Chlamydomonas reinhardtii*	[37]
	Plant	558	EST	Negative	*P. tremula–P. trichocarpa*	[20]
Protein–protein interaction	Yeast	164	Literatures, two-hybrid interactions	Negative	*S. cerevisiae–C. elegans*	[12]
	Yeast	13925 Interactions	Literatures, MIPS database	Negative	*S. cerevisiae–C. albicans*	[14]
	Yeast	~Total 3000 genes, 50000 interactions	MS data	NS[a]	*S. cerevisiae–C. albicans*	[33]

(continued on next page)

Table 2 (*continued*)

Genome variables	Species	# of total genes analyzed (NE/E)	Datasets for variables	Correlation	Species for divergence calculation (outgoup)	Ref.
	Yeast, bacteria	1004 Yeast, 500 Bacteria	MIPS database, PIMRider functional proteomics software platform	NS or weak negative	*S. cerevisiae*, *Schizosaccharomyces pombe*, *C. elgans*, two strains of *Helicobacter pylori* and *Campylobacter jejuni*	[61]
	Yeast, worm	4773 Yeast, 2386 nematode	DIP database	NS or weak negative	Six yeast species, 2 worms	[28]
	Yeast, worm, fly	20 252 Interactions	GRID database	Negative	*S. cerevisiae–S. paradoxus*, *C. elegans–Caenorhabditis griggsae*, *D. melanogaster–D. pseudoobscura*	[13]
	Fly	∼5000 Interactions	Two-hybrid interactions	Negative	*D. melanogaster–D. pseudoobscura* [illegible]	[15]
Essentially (effect of gene knockout)	Mouse	175 (108/67)	GKD database	NS	*M. musculus–R. norvegicus*	[7]
	Worm	19 213 Genes	RNAi phenotype/ microarray	Negative	*C. elegans–C. briggsae*	[11]
	Yeast	5724	Yeast deletion fitness data, *C. elegans*RNAi phenotype data	Negative	*S. cerevisiae*–9 yeast species	[19]
	Yeast	287 (119/168)	Yeast deletion fitness data	Negative	*S. cerevisiae–C. elegans*–5 bacterial and archaeal	[8]
	Yeast	3783	Yeast deletion fitness dataset	NS[a]	*S. cerevisiae*–3 yeast species and worms	[29]
	Yeast	1864	Yeast deletion fitness dataset	Negative	*S. cerevisiae–C. albicans*	[62]
	Yeast	3038	Yeast deletion fitness dataset	Negative	Four yeast species including *S. cerevisiae*	[10]
	Bacteria	1886 (1736/150)	PEC database	Negative	*E. coli–H. pylori–N. meningitidis* (2 strains for each)	[9]
	Bacteria	*B. subtilis*-4100 (?/277), *E. coli* ∼4300 (?/203)	*B. subtilis* deletion fitness dataset, PEC database	NS[a]	Five species including *E. coli* and *B. subtilis*	[36]

NE: non-essential genes, E: essential genes, NS: not significant, MS: mass spectrometry, Ref: reference, #: number.

[a] Correlation after expression abundance is controlled as a confounding factor.

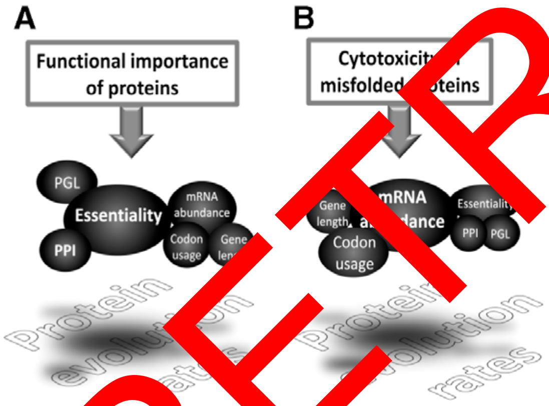

Fig. 1. Schematic drawings of the functional hypothesis (A) and the translational robustness hypothesis (B). (A) The functional hypothesis articulates that functional constraints dictate protein evolutionary rates. The hypothesis expects variables reflecting functional importance of proteins to correlate with protein evolutionary rates. Functional importance of a protein is supposed to determine the effect of gene knockout on the organism's fitness (named as 'essentiality' in this figure). In addition, PPI, PGL and mRNA abundance may associate with functional importance so that these variables would correlate with protein evolutionary rate. The size of each variable implies how much it reflects functional importance of proteins. The shadow indicates the contribution of each variable on determining protein evolutionary rate. (B) The MIM hypothesis suggests that the selection for a protein's robustness to lower mistranslation-induced misfolding should be particularly important for highly expressed proteins. Note that the expression abundance is the dominant correlate of the protein evolutionary rates. Other variables may associate with mRNA abundance resulting in their correlations with protein evolutionary rates. PGL: propensity of gene loss, PPI: protein–protein interaction.

or efficiency rather than for proper function of proteins may be critical determinant of the rates. One of the consequences of selection on efficient protein synthesis is co-adaptation of synonymous codon usage with tRNA pools. Among codons recognized by different aminoacyl tRNAs, translationally preferred codons tend to be recognized by more abundant isoacceptors. Protein abundance has been shown to correlate strongly with synonymous codon usage in some organisms [40]. Akashi has reported significantly higher frequency of preferred codons at conserved amino acids than at non-conserved ones in fruit flies [41]. Akashi [42] has also shown that in yeast there is a correlation between tRNA concentration and corresponding amino acid content that is stronger in highly expressed genes than in genes with low expression levels. Based on those findings, the translation accuracy hypothesis states that variations in the translation accuracy of different codons lead to selection of amino acids with better (or optimal) codons [42] and to counter-selecting non-synonymous changes leading to sub-optimal codons [43]; this in turn reduces the rate of protein evolution. Kimchi-Sarfaty et al. recently showed that synonymous mutations can contribute to a slow translation, thereby affecting the efficiency of cotranslational protein folding [44]. However, selection on codon usage is too weak to explain the slow pace of protein evolution since synonymous substitutions accumulate much faster than non-synonymous ones. In fact, when the better codons are removed from the analysis, the correlation between synonymous substitution rates and mRNA abundance disappears, but the association of mRNA abundance with non-synonymous substitution rates remains nearly unchanged [22]. How could the association between mRNA abundance (not protein abundance) and protein conservation be explained if not by translational accuracy?

5. Translational robustness hypothesis

The more frequently a protein is mistranslated and non-functional, the more the translational process costs. If the mistranslated proteins are toxic, it will have a greater fitness cost if it involves several proteins. Could typical frequencies of mistranslation (or ribosomal infidelity) be problematic for an organism? At an error rate of 5 × 10^{-4}, a 400-residue protein (an average length protein) can be expected to contain at least one mistranslation-derived missense mutation 18% of the time [21]. The incorporation of incorrect amino acids into proteins tends to destabilize them relative to the wild-type sequence, thus increasing their propensity to misfold. To reduce the number of proteins that misfold due to translation errors, selection can act on the amino acid sequence to increase the number of proteins that fold properly despite mistranslation. Hence, Drummond et al. [22] suggested that highly expressed proteins should be more tolerant to mistranslation. The authors called the increased tolerance for translational missense errors "translational robustness".

6. Mistranslation-induced protein misfolding hypothesis

Recently, Drummond and Wilke proposed the mistranslation-induced protein misfolding (MIM) hypothesis; adaption to reduce the cellular burden imposed by protein misfolding creates the prominent correlation between protein abundance and evolutionary rates (Table 1 and Fig. 1). The MIM hypothesis could explain the pervasive association of synonymous and non-synonymous substitution rates, since the cost of misfolded proteins can be reduced both at the translational level, by biasing codon usage to increase translational accuracy, and at the folding level, by favoring amino acid sequences with increased translational robustness. Using a molecular-level evolutionary simulation, Drummond and Wilke demonstrated that selection against toxicity of misfolded proteins generated by ribosome errors suffices to create all of the observed co-variation among genome variables [21]. The hypothesis is an attractive concept not only because it introduces a single, dominant determinant of protein evolutionary rate but also because the key role of translational robustness is compatible with fundamental biological features of all cells. Indeed, all cells encode numerous chaperones that prevent misfolding and enormously elaborate molecular machines such as proteasomes which to a large extent are dedicated to the selective degradation of misfolded proteins. Roughly 10–50% of random substitutions disrupt protein function [45,46]. Greater amounts of mistranslated protein may lead to elevated levels of toxic aggregates, especially if these mistranslated–misfolded proteins could seed the aggregation of the wild-type proteins by capturing folding intermediates [47,48]. More importantly, mistranslated proteins would definitely pose a burden on the proteostasis machinery of cells, leaving organisms more vulnerable to metabolic and environmental stresses [48] and less able to handle other [illegible] aggregation-prone proteins. Morimoto and colleagues have recently shown that the introduction of one protein prone to misfolding into a cell compromises that cell's ability to maintain proteostasis because other proteins begin to misfold and aggregate leading to proteotoxicity [49].

The burden of mistranslation-induced protein misfolding can be inferred by the association of misfolded proteins with several pathological conditions including neuronal degeneration, such as Alzheimer's disease, Huntington's disease, Parkinson's disease and amyotrophic lateral sclerosis [50]. Postmitotic neurons appear to be particularly sensitive to protein misfolding because aggregated toxic proteins cannot be diluted by cell division [51]. Malfunctioning of broadly expressed proteins involved in translation and protein folding manifests specifically neurotoxic effects in mouse [51,52] On the contrary, overexpression of chaperones has been reported to suppress neurodegeneration in fruit fly and mouse models [53,54]. Indeed, neurons were highlighted by Drummond and Wilke as being highly susceptible to translational infidelity and the fitness cost of misfolding [21]. Drummond and Wilke examined correlations between genome variables and tissue-specific mRNA levels in fly, mouse, and human, revealing that neural tissues have a stronger correlation of tissue expression with dN than do non-neural tissues [21].

Noticeably, the MIM hypothesis leads us to revisit the recent unusual finding made by Wyckoff et al., a positive correlation between dS and the dN/dS ratio [55]. Currently, no theory covers this observation, although Wyckoff et al. offer a possible explanation based on differences in mutation rates in different genes. The MIM hypothesis argues that a similar pressure against mistranslation would influence the evolution of both synonymous and non-synonymous substitution. If the selection generates greater variation in non-synonymous substitutions than in synonymous changes, dS would positively correlate with the dN/dS ratio.

The MIM hypothesis seems to make biological sense and explains data that the functional hypothesis hardly offers reasons for (see above). Although there is no experimental proof or confirmation of the hypothesis, Koonin and his colleagues recently reported insightful results. The dominant determinant of the sequence evolution rates is postulated to be the rate of translational events rather than mRNA or protein abundance. Given that the quantity that is actually measured in most experiments is the transcript level rather than the number of translation events per se, the interpretation of experimental data on gene expression is ambiguous. Avoiding the ambiguity, Koonin and his colleagues recently tested the hypothesis based on a simple yet elegant idea; different domains of the same protein are translated at the exact same rate [56]. They compared the evolutionary rates of 'individual domains fused into single proteins' against those of 'the same domains [illegible] into different proteins', and concluded that the translation rates are significant determinants of evolutionary rates. Nonetheless, definitive and conclusive experimental confirmation of the hypothesis is daunting since experimental measurement of translational robustness is challenging even in a handful of proteins, not to mention a genome-scale analysis.

7. Reconciliation of the MIM hypothesis with the pre-existed framework

Protein evolution requires two steps: the mutation of nucleotides that code for amino acids and the fixation of new variants in the population. The probability of fixation depends on the fitness effect of mutations; the new variant can be neutral or nearly neutral (and so governed purely or largely by genetic drift, respectively), deleterious (and consequently opposed by purifying selection), or advantageous (and therefore supported by positive selection) [57,58]. The MIM hypothesis suggests that there is a purifying selection against misfolded proteins, which results in a strong negative correlation between expression levels and protein evolutionary rates. Although the functional hypothesis may not be adequate to explain co-variation between the two variables, functional constraints may be a critical purifying selection pressure that lowers protein evolutionary rates in general. In other words, the premise of the MIM hypothesis is that two coding sequences under similar functional selective pressure might have differences in their evolutionary rates mostly due to other factors such as translational accuracy and translational robustness. In addition to the correlation of protein evolutionary rates with expression levels, it might be desirable to take into account other variables with small contributions to protein evolutionary rate for a more complete explanation [32,56].

There is now an increasing need to form a new integrated theory of protein evolution. We have both progressively sophisticated methods and genome-scale datasets to test individual evolutionary hypotheses that explain how genomic, cellular and physiological properties affect evolutionary process. An integrated view would combine these individual ideas and consider the global properties of proteins under a single conceptual framework. We anticipate that such a coherent theory will have far-reaching consequences on crucial problems in evolutionary biology. We believe that such a theory will require the integration of many individual elements including translational robustness.

Acknowledgments

We thank Bruce T. Lahn, Croydon Fernandes, and Seung Gu Park for critically reading the manuscript and helpful comments. This work was supported by the program of Nuclear R&D Program (2008-00903) of the Korea Science and Engineering Foundation (KOSEF).

References

[1] Margoliash, E. (1963) Primary structure and evolution of cytochrome *C*. Proc. Natl. Acad. Sci. USA 50, 672–679.

[2] Zuckerkandl, E. (1976) Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. J. Mol. Evol. 7, 167–183.

[3] Koonin, E.V. and Wolf, Y.I. (2006) Evolutionary systems biology: links between gene evolution and function. Curr. Opin. Biotechnol. 17, 481–487.

[4] Wolf, Y.I., Carmel, L. and Koonin, E.V. (2006) Unifying measures of gene function and evolution. Proc. Biol. Sci. 273, 1507–1515.

[5] Koonin, E.V. (2005) Systemic determinants of gene evolution and function. Mol. Syst. Biol. 1, 0021.

[6] Drummond, D.A., Raval, A. and Wilke, CO. (2006) A single determinant dominates the rate of yeast protein evolution. Mol. Biol. Evol. 23, 327–337.

[7] Hurst, LD. and Smith, N.G. (1999) Do essential genes evolve slowly? Curr. Biol. 9, 747–750.

[8] Hirsh, A.E. and Fraser, H.B. (2001) Protein dispensability and rate of evolution. Nature 411, 1046–1049.

[9] Jordan, I.K., Rogozin, I.B., Wolf, Y.I. and Koonin, E.V. (2002) Essential genes are more evolutionary conserved than are nonessential genes in bacteria. Genome Res. 12, 962–968.

[10] Wall, D.P., Hirsh, A.E., Fraser, H.B., Kumm, J., Giaever, G., Eisen, M.B. and Feldman, M.W. (2005) Functional genomic analysis of the rates of protein evolution. Proc. Natl. Acad. Sci. USA 102, 5483–5488.

[11] Cutter, A.D. et al. (2003) Molecular correlates of genes exhibiting RNAi phenotypes in *Caenorhabditis elegans*. Genome Res. 13, 2651–2657.

[12] Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C. and Feldman, M.W. (2002) Evolutionary rate in the protein interaction network. Science 296, 750–752.

[13] Hahn, M.W. and Kern, A.D. (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol. Biol. Evol. 22, 803–806.

[14] Fraser, H.B., Wall, DP. and Hirsh, A.E. (2003) A simple dependence between protein evolution rate and the number of protein–protein interactions. BMC Evol. Biol. 3, 11.

[15] Lemos, B., Bettencourt, B.R., Meiklejohn, C.D. and Hartl, D.L. (2005) Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein–protein interactions. Mol. Biol. Evol. 22, 1345–1354.

[16] Mintseris, J. and Weng, Z. (2005) Structure, function, and evolution of transient and obligate protein–protein interactions. Proc. Natl. Acad. Sci. USA 102, 10930–10935.

[17] Pal, C., Papp, B. and Hurst, L.D. (2001) Highly expressed genes in yeast evolve slowly. Genetics 158, 927–931.

[18] Subramanian, S. and Kumar, S. (2004) Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 168, 373–381.

[19] Zhang, J. and He, X. (2005) Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol. Biol. Evol. 22, 1147–1155.

[20] Ingvarsson, P.K. (2007) Gene expression and protein length influence codon usage and rates of sequence evolution in *Populus tremula*. Mol. Biol. Evol. 24, 836–844.

[21] Drummond, D.A. and Wilke, CO. (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134, 341–352.

[22] Drummond, D.A., Bloom, J.D., Adami, C., Wilke, CO. and Arnold, F.H. (2005) Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. USA 102, 14338–14343.

[23] Duret, L. and Mouchiroud, D. (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. 17, 68–74.

[24] Krylov, D.M., Wolf, Y.I., Rogozin, I.B. and Koonin, E.V. (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. 13, 2229–2235.

[25] Batada, N.N., Hurst, L.D. and Tyers, M. (2006) Evolutionary and physiological importance of hub proteins. PLoS Comput. Biol. 2, e88.

[26] Batada, N.N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.J., Hurst, L.D. and Tyers, M. (2007) Still stratus not altocumulus: further evidence against the date/party hub distinction. PLoS Biol. 5, e154.

[27] Bertin, N., Simonis, N., Dupuy, D., Cusick, M.E., Han, J.D., Fraser, H.B., Roth, F.P. and Vidal, M. (2007) Confirmation of organized modularity in the yeast interactome. PLoS Biol. 5, e153.

[28] Agrafioti, I., Swire, J., Abbott, J., Huntley, D., Butcher, S. and Stumpf, M.P. (2005) Comparative analysis of the *Saccharomyces cerevisiae* and *Caenorhabditis elegans* protein interaction networks. BMC Evol. Biol. 5, 23.

[29] Pal, C., Papp, B. and Hurst, L.D. (2003) Genomic function: rate of evolution and gene dispensability. Nature 421, 496–497 [discussion 497–498].

[30] Bloom, J.D. and Adami, C. (2004) Evolutionary rate depends on number of protein–protein interactions independently of gene expression level: response. BMC Evol. Biol. 4, 14.

[31] Plotkin, J.B. and Fraser, H.B. (2007) Assessing the determinants of evolutionary rates in the presence of noise. Mol. Biol. Evol. 24, 1113–1121.

[32] Bloom, J.D., Drummond, D.A., Arnold, F.H. and Wilke, C.O. (2006) Structural determinants of the rate of protein evolution in yeast. Mol. Biol. Evol. 23, 1751–1761.

[33] Bloom, J.D. and Adami, C. (2003) 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.

[34] Mclnerney, J.O. (2006) The causes of protein evolutionary rate variation. Trends Ecol. Evol. 21, 230–232.

[35] Rocha, E.P. (2006) The quest for the universals of protein evolution. Trends Genet. 22, 412–416.

[36] Rocha, E.P. and Danchin, A. (2004) An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol. Biol. Evol. 21, 108–116.

[37] Popescu, C.E., Borza, T., Bielawski, J.P. and Lee, R.W. (2006) Evolutionary rates and expression level in *Chlamydomonas*. Genetics 172, 1567–1576.

[38] Wright, S.I., Yau, C.B., Looseley, M. and Meyers, B.C. (2004) Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. Mol. Biol. Evol. 21, 1719–1726.

[39] Marais, G. and Duret, L. (2001) Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. J. Mol. Evol. 52, 275–280.

[40] Gouy, M. and Gautier, C. (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10, 7055–7074.

[41] Akashi, H. (1994) Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics 136, 927–935.

[42] Akashi, H. (2003) Translational selection and yeast proteome evolution. Genetics 164, 1291–1303.

[43] Lipman, D.J. and Wilbur, W.J. (1984) Interaction of silent and replacement changes in eukaryotic coding sequences. J. Mol. Evol. 21, 161–167.

[44] Kimchi-Sarfaty, C., Oh, J.M., Kim, I.W., Sauna, Z.E., Calcagno, A.M., Ambudkar, S.V. and Gottesman, M.M. (2007) A silent polymorphism in the MDR1 gene changes substrate specificity. Science 315, 525–528.

[45] Guo, H.H., Choe, J. and Loeb, L.A. (2004) Protein tolerance to random amino acid change. Proc. Natl. Acad. Sci. USA 101, 9205–9210.

[46] Markiewicz, P., Kleina, L.G., Cruz, C., Ehret, S. and Miller, J.H. (1994) Genetic studies of the lac repressor. XIV. Analysis of 4000 altered *Escherichia coli* lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. J. Mol. Biol. 240, 421–433.

[47] Balch, W.E., Morimoto, R.I., Dillin, A. and Kelly, J.W. (2008) Adapting proteostasis for disease intervention. Science 319, 916–919.

[48] Cohen, E., Bieschke, J., Perciavalle, R.M., Kelly, J.W. and Dillin, A. (2006) Opposing activities protect against age-onset proteotoxicity. Science 313, 1604–1610.

[49] Gidalevitz, T., Ben-Zvi, A., Ho, K.H., Brignull, H.R. and Morimoto, R.I. (2006) Progressive disruption of cellular protein folding in models of polyglutamine diseases. Science 311, 1471–1474.

[50] Soto, C. (2003) Unfolding the role of protein misfolding in neurodegenerative diseases. Nat. Rev. Neurosci. 4, 49–60.

[51] Lee, J.W. et al. (2006) Editing-defective tRNA synthetase causes protein misfolding and neurodegeneration. Nature 443, 50–55.

[52] Zhao, L., Longo-Guess, C., Harris, B.S., Lee, J.W. and Ackerman, S.L. (2005) Protein accumulation and neurodegeneration in the woozy mutant mouse is caused by disruption of SIL1, a cochaperone of BiP. Nat. Genet. 37, 974–979.

[53] Cummings, C.J., Sun, Y., Opal, P., Antalffy, B., Mestril, R., Orr, H.T., Dillmann, W.H. and Zoghbi, H.Y. (2001) Over-expression of inducible HSP70 chaperone suppresses neuropathology and improves motor function in SCA1 mice. Hum. Mol. Genet. 10, 1511–1518.

[54] Auluck, P.K., Chan, H.Y., Trojanowski, J.Q., Lee, V.M. and Bonini, N.M. (2002) Chaperone suppression of alpha-synuclein toxicity in a Drosophila model for Parkinson's disease. Science 295, 865–868.

[55] Wyckoff, G.J., Malcom, CM., Vallender, E.J. and Lahn, B.T. (2005) A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. Trends Genet. 21, 381–385.

[56] Wolf, M.Y., Wolf, Y.I. and Koonin, E.V. (2008) Comparable contributions of structural–functional constraints and expression level to the rate of protein sequence evolution. Biol. Direct. 3, 40.

[57] Pal, C., Papp, B. and Lercher, M.J. (2006) An integrated view of protein evolution. Nat. Rev. Genet. 7, 337–348.

[58] Bamshad, M. and Wooding, S.P. (2003) Signatures of natural selection in the human genome. Nat. Rev. Genet. 4, 99–111.

[59] Zhou, T., Drummond, D.A. and Wilke, CO. (2008) Contact density affects protein evolutionary rate from bacteria to animals. J. Mol. Evol. 66, 395–404.

[60] Jordan, I.K., Marino-Ramirez, L., Wolf, Y.I. and Koonin, E.V. (2004) Conservation and coevolution in the scale-free human gene coexpression network. Mol. Biol. Evol. 21, 2058–2070.

[61] Jordan, I.K., Wolf, Y.I. and Koonin, E.V. (2003) No simple dependence between protein evolution rate and the number of protein–protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol. Biol. 3, 1.

[62] Yang, J., Gu, Z. and Li, W.H. (2003) Rate of protein evolution versus fitness effect of gene deletion. Mol. Biol. Evol. 20, 772–774.