Published in final edited form as:

Curr Opin Struct Biol. 2010 June ; 20(3): 385–389. doi:10.1016/j.sbi.2010.03.004.

Signatures of protein biophysics in coding sequence evolution

Claus O. Wilke^{1,*} and D. Allan Drummond²

¹ Center for Computational Biology and Bioinformatics, Institute for Cell and Molecular Biology, and Section of Integrative Biology, The University of Texas at Austin, Austin, TX.

² FAS Center for Systems Biology, Harvard University, Boston, MA.

Abstract

Since the early days of molecular evolution, the conventional wisdom has been that the evolution of protein-coding genes is primarily determined by functional constraints. Yet recent evidence indicates that the evolution of these genes is strongly shaped by the biophysical processes of protein synthesis, protein folding, and specific as well as non-specific protein–protein interactions. Selection pressures related to these biophysical processes affect primarily the amino-acid sequence of genes, but they also leave their mark on synonymous sites at the nucleotide level. While evidence for specific selection pressures related to protein biophysics is strong, there is currently no unifying framework that integrates the various selection pressures on coding sequences and disentangles their relative importance.

1 Introduction

With the availability of hundreds of fully-sequenced genomes, the comparative analysis of genomes and, more specifically, of protein-coding genes has become commonplace. One striking observation in such comparisons is that sequence conservation varies dramatically among different genes in the same species. Consider, for example, a comparison of mouse and human. Some pairs of orthologs in these two species have hardly changed at all while others have changed so much that we can barely tell that they are orthologs. In general, the rate with which genes accumulate substitutions over time varies over several orders of magnitude and is lognormally distributed [1-3]. This result is fairly independent of the specific organisms under study. The challenge, thus, is to explain why some genes evolve so much faster than others. Recent work has indicated that the answer to this question will likely involve protein biophysics, including protein synthesis, protein folding, and protein–protein interactions.

Protein-coding genes evolve through amino-acid substitutions, through insertion or deletion of individual amino acids, and through the addition, deletion, or rearrangement of entire protein domains. Here, we focus exclusively on patterns of amino-acid substitutions. We review recent advances in our understanding of how protein biophysics shapes the accumulation of substitutions at the level of the gene.

© 2010 Elsevier Ltd. All rights reserved.

* Corresponding author cwilke@mail.utexas.edu.
dadrummond@cgr.harvard.edu

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Patterns of amino-acid substitutions are usually assessed by measures of sequence conservation. As a general rule, sequence conservation equals constraint. The more conserved a sequence region is over evolutionary time, the more selective constraint it experiences (purifying selection). Sites that are strongly constrained experience few amino-acid substitutions over evolutionary time, whereas sites under little constraint accumulate substitutions much faster. For multiple sites or entire proteins, the conventional measure of sequence conservation is the *evolutionary rate*. It is defined as the mean rate at which the sites under consideration accumulate mutations over evolutionary time. In this review, unless specified otherwise, we focus on non-synonymous or amino-acid changing evolution. Thus, we use the term “evolutionary rate” interchangeably with the more precise term “non-synonymous evolutionary rate.”

By definition, evolutionary rate is an average. If we calculate the evolutionary rate over ten sites, of which nine evolve slowly and one evolves fast, then the overall rate for the ten sites will be low. Thus, to first order, evolutionary rates measure how many sites in a gene are not strongly constrained. Low evolutionary rates mean most sites are under purifying selection, and high evolutionary rates mean only a few sites are under purifying selection.

2 Protein structure shapes sequence evolution

Most proteins need to fold stably into their correct conformation in order to function. Therefore, mutations that impair either the stability or the conformation of the folded protein will generally be selected against. Different sites in a protein differ in the extent to which they admit changes in protein stability or fold upon mutation. The more sensitive sites show reduced evolutionary variation while the more robust sites are more variable.

The primary predictor of whether a specific site is robust or sensitive is the relative solvent accessibility of the residue at that site [4-6]. Buried sites evolve about half as rapidly as exposed sites [4]. The secondary structure at a given site, on the other hand, has comparatively little influence on how fast a site evolves [4]. See Ref. [7] for a recent review on this topic.

Since solvent-exposed residues are more variable than buried residues, we might expect that proteins with a larger fraction of solvent-exposed residues evolve faster overall. Yet the opposite is the case: The fraction of buried sites correlates positively with the overall evolutionary rate of the protein, and explains 5% to 10% of the variation in evolutionary rate [6,8,9]. (Ref. [10] found the opposite result. They found this result, we believe, because they predicted solvent accessibilities using a support-vector machine rather than measuring them on actual protein structures. See discussion in [9].) This finding seems paradoxical. The solution to the paradox was hinted at by Bloom et al. [8] and shown clearly in an elegant recent analysis by Franzosa and Xia [6]. A large protein core tends to stabilize the overall protein and allows the surface residues to vary more freely. The increased variability of the surface residues more than compensates for the reduction in the relative number of surface residues in proteins with a large core. This reasoning is consistent with theoretical predictions on the number of sequences compatible with a given structure [11].

Many proteins contain regions that rarely if ever adopt a well-defined three-dimensional structure. These natively disordered regions may provide flexible links between domains, fold only in the presence of a ligand, or carry out a function that does not require a well-defined structure. As would be predicted from the hypothesis that the requirement for structure imposes an evolutionary constraint on protein sequences, natively disordered regions evolve more rapidly than structured regions [12,13].

3 Translational selection

One of the best predictors of evolutionary rate is expression level, measured by mRNA or protein abundance—the more highly expressed a gene is, the slower it evolves. This relationship has been found in nearly every organism for which evolutionary rates and expression measurements are available [2,14,15]. Many selection pressures hypothesized to act on protein sequences would be amplified by expression level (Figure 1). For example, the total cost of synthesizing all the protein expressed from a given gene increases with the gene's expression level, and thus the effect of a mutation that changes the per-protein synthesis cost will grow in proportion to expression level, all else equal. The same reasoning applies to costs related to protein degradation and protein toxicity, and indirect costs linked to the production of other proteins, such as chaperones, that interact directly with the protein of interest.

Highly expressed proteins evolve slowly even when considering only evolutionary-rate differences between duplicate genes [16]. Because these genes tend to share both structure and biological function, the explanation for the expression–rate relationship is unlikely to involve either feature. Accordingly, expression level has been hypothesized to constrain protein sequences due to selection against production of misfolded proteins. First, errors during translation, such as missense substitutions, can disrupt protein folding, leading to selection for coding sequences that translate with reduced error rates (increased accuracy, see also section “Selection on synonymous sites”) and encode proteins with increased tolerance for errors (increased robustness) [2,16]. Improved robustness could be obtained by small increases in thermodynamic stability [2,16,17]. Second, selection may favor protein sequences that fold with high reliability, e.g., with improved kinetic folding properties that reduce misfolding of error-free sequences [2,18].

These adaptations are hypothesized to slow the rate of substitution because accuracy, robustness, and reliability are all rare properties, that is, given increasing need for these properties in order to avoid misfolding, the set of sequences having the required properties shrinks substantially. If most substitutions contributing to evolutionary-rate differences accumulate due to random drift among a set of sequences compatible with constraints on protein misfolding, then selection on these adaptations will slow the rate of sequence evolution.

Experimental tests of these hypotheses must still be done. However, some experimental evidence consistent with a key prediction of the translational-robustness hypothesis has accumulated. When β lactamase is expressed by an error-prone RNA polymerase, it evolves increased thermostability [19]. In addition, a recent theoretical study concluded that highly expressed and slowly evolving proteins share some sequence characteristics with thermostable proteins [20].

4 Protein–protein interactions

Most proteins interact with other proteins to carry out their function. These interactions can be obligate (when individual proteins are part of a larger complex, such as the ribosome) or transient (when proteins dock to other proteins for short periods of time or under special conditions, such as in signal transduction). For both kinds of interactions, the residues in the interface need to be carefully chosen for a successful interaction to be possible. Therefore, protein–protein interfaces should be more strongly conserved than regions of protein surfaces that are non-interacting. This hypothesis is generally supported by the available data [6,21–24].

Because protein–protein interactions create additional constraints on protein sequences, some authors have argued that the overall evolutionary rate of a protein should depend on the number of interaction partners that a protein has [25–27]. Yet this signal is rather weak, and much of it appears to be caused by confounding variables [28–31]. There are multiple reasons why overall evolutionary rate may not be much affected by the number of interaction partners. First, because evolutionary rate is an average over all sites in a protein, the *fraction* of sites involved in interactions should be more important than the *number* of interaction partners [23]. Second, obligate interactions seem to constrain interfaces more strongly than transient interactions do [22], and these differences in the amount of constraint are not accounted for when simply counting the number of interaction partners. Third, it seems inconsistent to only consider specific interactions but to disregard non-specific interactions. All proteins are at risk of participating in random, non-specific interactions that will generally be deleterious [32,33]. Therefore, non-interface surface regions must experience selection pressure to avoid unintended protein–protein interactions. This selection pressure will likely increase with gene expression level, which may explain why the differences in sequence conservation between interface and non-interface regions decrease with increasing expression level [24].

5 Selection on synonymous sites

Selection for reliable and accurate protein folding shapes evolution at the amino-acid level, as discussed in the previous sections. Surprisingly, the same selection pressures have measurable influence on synonymous (“silent”) substitutions at the nucleotide level. This influence arises through the interaction between the mRNA, the ribosome, and the nascent polypeptide. There are two main mechanisms. First, codon choice can affect the speed of translation; some codons—those corresponding to highly abundant tRNAs—are translated faster than others. Therefore, synonymous codon choice can cause the ribosome to stall or to speed up [34,35]. The change in ribosome speed may interact with the co-translational folding of the nascent polypeptide. Several examples are known where precisely regulated ribosome speed, mediated by codon choice, is believed to be necessary for accurate folding of the synthesized protein [34–37].

Second, codon choice can affect the accuracy of translation. Different synonymous codons have different error rates [38]. In general, codons with lower error rate will have a selective advantage over synonymous codons with higher error rate because the former provide greater protection from translation errors. However, at any given site, the strength of this selection pressure depends on the specific function of the residue at that site. For example, a residue that is located in the core of the protein and is critical for proper folding would have to be protected from translation errors, whereas a residue that is located in a loop and has no specific function might be quite tolerant to translation errors and would not require protection. The signal to indicate translational-accuracy selection is thus a statistical association between preferred codons and structurally or functionally important sites. Such an association has been found by several authors, who used as measure of a site's importance evolutionary conservation [2,39,40], involvement in a specific function [39], or relative solvent accessibility in the folded protein [41,42].

Other selection pressures on synonymous sites arise in the context of translation initiation, splicing, or nucleosome positioning [43–45].

6 Relative importance of different effects

Many different constraints on protein evolution have been hypothesized, and some are reasonably well-understood, as reviewed in the preceding sections. What is much less clear, however, is how these mechanisms interact and what their relative contributions are to the

evolutionary dynamics of a given protein. Two recent works have started to address this question.

Franzosa and Xia asked to what extent sequence conservation at protein interfaces is due to reduced solvent accessibility versus the requirement to achieve accurate and reliable docking [6]. When two proteins dock, the solvent accessibility of the interface residues is reduced. This reduction in solvent accessibility by itself could be responsible for increased sequence conservation at interface sites. Franzosa and Xia compared the observed sequence conservation of interface residues to their expected value given the reduction in solvent accessibility upon complex formation [6]. They found that interfaces generally imposed an additional selection pressure beyond just what would be expected from reduced solvent accessibility.

Wolf et al. [46] addressed the question to what extent gene expression level constrains protein evolution relative to intrinsic features of a protein such as its structure or function. They compared the evolution of domains in multi-domain proteins to the evolution of homologous domains in other proteins. Two domains in a multi-domain protein experience exactly the same expression level. Therefore, if expression level alone determined evolutionary rate, these two domains would evolve at exactly the same rate. By contrast, if intrinsic features dominated, each domain would evolve at its own individual rate. Wolf et al. found that both expression level and intrinsic factors are important, and that they have effects of comparable magnitude [46].

7 Conclusions

The molecular evolution of protein-coding genes takes place in the context of protein structure and protein folding. Over the last few years, we have witnessed a rapid growth in a body of research analyzing gene evolution from a biophysical perspective. This work has shown that biophysical processes such as protein synthesis and protein folding are among the most important factors shaping coding-sequence evolution. Despite the rapid progress in this field, we believe that important questions remain unanswered. In particular, we believe that the next major challenge will be to assess the relative importance of various mechanisms. Some authors have made first steps in this direction, but a systematic and comprehensive approach is currently lacking. The development of such an approach will be a major contribution to the field of molecular evolution.

Acknowledgments

This work was supported by NIH grants P50 GM068763 and R01 GM088344.

References

1. Grishin NV, Wolf YI, Koonin EV. From complete genomes to measures of substitution rate variability within and between proteins. *Genome Res* 2000;10:991–1000. [PubMed: 10899148]
- 2*. Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 2008;134:341–352. Patterns of evolutionary-rate variation are conserved from bacteria to mammals and can be reproduced by a model of mistranslation-induced protein misfolding. [PubMed: 18662548]
3. Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci USA* 2009;106:7273–7280. [PubMed: 19351897]
4. Goldman N, Thorne J, Jones D. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 1998;149:445–458. [PubMed: 9584116]

5. Mirny LA, Shakhnovich EI. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol* 1999;291:177–196. [PubMed: 10438614]
- 6*. Franzosa EA, Xia Y. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol* 2009;26:2387–2395. Interacting residues on protein surfaces experience a stronger reduction in their evolutionary rate than expected from the reduction in solvent accessibility caused by the interaction. [PubMed: 19597162]
7. Worth CL, Gong S, Blundell TL. Structural and functional constraints in the evolution of protein families. *Nature Reviews Mol Cell Biol* 2009;10:709–719.
8. 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]
9. Zhou T, Drummond DA, Wilke CO. Contact density affects protein evolutionary rate from bacteria to animals. *J Mol Evol* 2008;66:395–404. [PubMed: 18379715]
10. Lin YS, Hsu WL, Hwang JK, Li WH. Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Mol Biol Evol* 2007;24:1005–1011. [PubMed: 17264066]
11. England JL, Shakhnovich EI. Structural determinant of protein designability. *Phys Rev Lett* 2003;90:218101. [PubMed: 12786593]
12. Brown C, Takayama S, Campen A, Vise P, Marshall T, Oldfield C, Williams C, Dunker A. Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol* 2002;55:104–110. [PubMed: 12165847]
13. Brown C, Johnson A, Daughdrill G. Comparing models of evolution for ordered and disordered proteins. *Mol Biol Evol*. 2009 in press.
14. 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]
15. Lemos B, C D Meiklejohn BRB, Hartl DL. 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* 2005;22:1345–1354. [PubMed: 15746013]
16. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA* 2005;102:14338–14343. [PubMed: 16176987]
17. Wilke CO, Drummond DA. Population genetics of translational robustness. *Genetics* 2006;173:473–481. [PubMed: 16489231]
18. Drummond DA, Wilke CO. The evolutionary consequences of erroneous protein synthesis. *Nature Reviews Genetics* 2009;10:715–724.
- 19*. Goldsmith M, Tawfik DS. Potential role of phenotypic mutations in the evolution of protein expression and stability. *Proc Natl Acad Sci USA* 2009;106:6197–6202. TEM 1 β -lactamase adapts to an increased error frequency under transcription by evolving increased thermodynamic stability. [PubMed: 19339491]
20. Cherry JL. Highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins. *Mol Biol Evol*. 2010 in press.
21. Valdar WSJ, Thornton JM. Protein–protein interfaces: Analysis of amino-acid conservation in homodimers. *PROTEINS: Structure, Function, and Genetics* 2001;42:108–124.
22. Mintseris J, Weng Z. Structure, function, and evolution of transient and obligate protein–protein interactions. *Proc Natl Acad Sci USA* 2005;102:10930–10935. [PubMed: 16043700]
23. Kim PM, Lu LJ, Gerstein MB. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* 2006;314:1882–1883. [PubMed: 17185587]
24. Eames M, Kortemme T. Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mRNA level and protein abundance. *Structure* 2007;15:1442–1451. [PubMed: 17997970]
25. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. *Science* 2002;296:750–752. [PubMed: 11976460]
26. Fraser HB, Wall DP, Hirsh AE. A simple dependence between protein evolution rate and the number of protein-protein interactions. *BMC Evol Biol* 2003;3:11. [PubMed: 12769820]

27. Fraser H, Plotkin J. Assessing the determinants of evolutionary rates in the presence of noise. *Mol Biol Evol* 2007;24:1113–1121. [PubMed: 17347158]
28. 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]
29. 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]
30. Xia Y, Franzosa EA, Gerstein MB. Integrated assessment of genomic correlates of protein evolutionary rate. *PLoS Comp Biol* 2009;5:e1000413.
31. Wolf YI, Carmel L, Koonin EV. Unifying measures of gene function and evolution. *Proc R Soc B* 2006;273:1507–1515.
32. Deeds EJ, Ashenberg O, Gerardin J, Shakhnovich EI. Robust protein–protein interactions in crowded cellular environments. *Proc Natl Acad Sci USA* 2007;104:14952–14957. [PubMed: 17848524]
- 33*. Zhang J, Maslov S, Shakhnovich EI. Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size. *Mol Syst Biol* 2009;4:210. Non-specific protein–protein interactions limit proteome diversity and gene expression levels. [PubMed: 18682700]
34. Komar AA, Lesnik T, Reiss C. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett* 1999;462:387–391. [PubMed: 10622731]
35. Zhang G, Hubalewska M, Ignatova Z. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature Struct Mol Biol* 2009;16:274–280. [PubMed: 19198590]
36. Cortazzo P, Cerveñansky C, Marín M, Reiss C, Ehrlich R, Deana A. Silent mutations affect in vivo protein folding in *Escherichia coli*. *Biochem Biophys Res Comm* 2002;293:537–541. [PubMed: 12054634]
37. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM. A “silent” polymorphism in the *mdr1* gene changes substrate specificity. *Science* 2007;315:525–528. [PubMed: 17185560]
38. Kramer EB, Farabaugh PJ. The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. *RNA* 2007;13:87–96. [PubMed: 17095544]
39. Akashi H. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* 1994;136:927–935. [PubMed: 8005445]
40. Stoletzki N, Eyre-Walker A. Synonymous codon usage in *Escherichia coli*: Selection for translational accuracy. *Mol Biol Evol* 2007;24:374–381. [PubMed: 17101719]
- 41*. Zhou T, Weems M, Wilke CO. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol* 2009;26:1571–1580. Preferred codons associate with buried sites in proteins. [PubMed: 19349643]
42. Warnecke T, Hurst LD. GroEL dependency affects codon usage—support for a critical role of misfolding in gene evolution. *Mol Syst Biol* 2010;6:340. [PubMed: 20087338]
43. Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 2009;324:255–258. [PubMed: 19359587]
44. Gu W, Zhou T, Wilke CO. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comp Biol*. 2010 in press.
45. Warnecke T, Weber CC, Hurst LD. Why there is more to protein evolution than protein function: splicing, nucleosomes and dual-coding sequence. *Biochem Soc Transact* 2009;37:756–761.
- 46*. Wolf MY, Wolf YI, Koonin EV. Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution. *Biology Direct* 2008;3:40. The evolutionary rate of protein domains is determined in part by gene expression level and in part by intrinsic structural and functional constraints on the domain. [PubMed: 18840284]
47. Willensdorfer M, Bürger R, Nowak MA. Phenotypic mutation rates and the abundance of abnormal proteins in yeast. *PLoS Comput Biol* 2007;3:e203. [PubMed: 18039025]

Figure 1.

Expression level amplifies specific fitness costs. (A) Fitness costs related to protein synthesis, misfolding, aggregation, or non-specific interactions generally grow with increasing expression level, all else being equal [33,47]. By contrast, fitness costs related to function do not necessarily grow in proportion to expression level. For example, genes with both high and low fitness cost upon knockout can be found at all expression levels (B). The data in (B) are for *Saccharomyces cerevisiae*, taken from [29].