Update TRENDS in Biotechnology Vol.23 No.10 October 2005

Research Focus

Converging on a general model of protein evolution

Joshua T. Herbeck[1] and Dennis P. Wall[2]

[1]Department of Microbiology, University of Washington School of Medicine, Seattle, WA 98103, USA
[2]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA

The availability of high-throughput genomic databases that establish protein dispensability, expression and interaction networks enables rigorous tests of competing models of protein evolution. Recent research utilizing these new data sets shows that protein evolution is more complex than was previously thought. Several variables, including protein dispensability, expression, functional density, and genetic modularity, appear to have independent effects on the evolutionary rate of proteins, suggesting that proteomes have evolved via an assembly of selectional regimes. These results indicate that a general model of protein evolution will emerge as more functional genomic data from a diversity of organisms accumulate.

Introduction

It has been known for decades that the evolutionary rates of proteins vary by several orders of magnitude, but the causes of this variation are only beginning to be understood more clearly. The neutral theory of molecular evolution [1] predicts that the rate of protein evolution depends, at least in part, on the overall importance of the protein to the fitness of the organism. This proposition can be separated into two hypotheses. First, in less important (*i.e.* more dispensable) proteins, a larger fraction of the mutations will have neutral effect on fitness and, thus, should arise unimpeded by natural selection. Second, the protein evolutionary rate will correlate with the density of amino acids that have high functional importance, with proteins in which this density is greater evolving more slowly than those in which it is lesser.

The factors that are hypothesized to influence protein evolutionary rate (Table 1) are difficult to test because the measurement of appropriate variables is exigent. Furthermore, the extent of the experimental noise in these data is difficult to determine with any degree of certainty. Thus, it is not surprising that a series of conflicting opinions on the relative effects of different functional genomic variables on protein evolutionary rate has arisen since the first attempts to test these hypotheses [2–7]. As a consequence of this debate, a third hypothesis has been proposed in which protein evolutionary rate depends on neither dispensability nor functional density in any appreciable way, but on the rate of protein expression. This prediction stems from the hypothesis that proteins are under selectional constraint for accurate translation and metabolic efficiency [8,9]. Indeed, in some studies dispensability and functional density are largely covariates of protein expression rate, and controlling for protein expression rate causes the apparent relationships between dispensability and evolutionary rate, and between functional density and evolutionary rate, to disappear [6]. However, recent improvements in speed, cost efficiency and scope of the protocols used to gather functional and comparative genomic data (Table 2) have enabled more robust analyses of protein evolution.

Dispensability and expression both influence protein evolutionary rate

Wall *et al.* [10] have used some of the best-available functional genomic and evolutionary data to test the independent effects of dispensability and expression on evolutionary rate. Their analysis is strengthened by an expanded, comparative genomic database that consists of four, closely related species of yeast [11], and by a much larger data set of dispensability [12]. They show that both dispensability and expression have significant, but independent, effects on the rate of protein evolution. This result is confirmed in a similar study by Zhang and He [13]. Zhang and He also measured correlations between protein dispensability and evolutionary rate using rate data calculated from a larger comparative genomic data set that includes *Saccharomyces cerevisiae* and nine other yeast species. They made the crucial observation that the effect of gene expression on evolutionary rate is weaker in comparisons of more closely related species. In effect, the

Table 1. Factors hypothesized to influence protein evolutionary rate

Protein variable	Synopsis of hypothesis	Notable tests
Dispensability	The relative dispensability of a protein to the overall fitness of an organism influences the evolutionary rate because more dispensable proteins have more sites that are selectively neutral and subject to genetic drift	[1,15,24–26]
Functional-density	The rate of evolution of a protein depends on how many of its sites are constrained by function	[4,5,15,16]
Expression	The rate of expression of a protein determines its evolutionary rate because of differential selection for accurate translation and/or metabolic efficiency	[8,9,27]
Modularity	The degree of connectivity and nature of connectivity of a protein in the protein-interaction network influences its evolutionary rate	[20,21]

Corresponding author: Herbeck, J.T. (herbeck@u.washington.edu).
Available online 27 July 2005

Table 2. High-throughput methods for collecting functional genomic variables to test the hypotheses in Table 1

Protein variable	Example of high-throughput measurement protocol	Refs
Dispensability	Growth rate of knockout strains. In yeast, each gene is deleted from a single strain, and all deletion mutants are grown in a common medium. The rates of growth are measured by fluorescence intensities	[12,22,28,29]
Functional density	The number of protein–protein interactors per protein. Protein–protein interactions are measured by two-hybrid screens, mass spectrometry, synthetic lethality, computational prediction and other approaches	[19,20,30–33]
Expression	Rate of transcription, rate of translation, mRNA abundance, protein abundance and level of synonymous codon usage bias	[10,14,34,35]
Evolutionary rate	Rate of nonsynonymous and synonymous mutations, estimated by either pairwise or phylogeny-based comparison of orthologous gene sequences	[10,13,23,36]
Genetic modularity	The delimitation of protein–protein interaction networks into compartments of proteins that perform similar functions	[20,21]

availability of high-throughput genomic data was the key factor that enabled a more robust analysis of variables that affect protein evolution.

In addition to the improved dispensability data and evolutionary-rate estimations in these studies, Wall *et al.* [10] account for noise in two important ways. In the first case, they used a structural equation model to partition the variance in the data caused by experimental error from the variance caused by real biological effects. Their observed correlations (between dispensability, expression and evolutionary rate) are consistent with a range of correlations observed under variable levels of noise in the data. In the second case, Wall *et al.* account for noise in the data by cross-referencing functional genomic databases. A significant but, at present, unavoidable source of noise in many high-throughput databases is the laboratory conditions under which they are measured. Wall *et al.* propose that the difference between the laboratory and the natural environment is likely to be less significant in genes that are expressed more highly. They used array-based gene-expression data [14] to filter the dispensability data accordingly and, in doing so, significantly improved the strengths of the correlations. This heuristic approach should be applicable to many fields of study that use similar, high-throughput genomic databases.

Protein-interaction networks and protein evolutionary rate

In addition to dispensability and expression, another variable that is thought to influence the rate of evolution of a protein is the density of functionally important sites in that protein [15]. Early studies of the structure and function of individual proteins indicated that molecular interactions impose constraints on protein evolution because they require precisely organized structures [16]. Experimental advances in the characterization of protein-interaction networks have begun to make tests of this hypothesis possible [17]. Fraser and colleagues first tested [18] and, subsequently, confirmed [4] that the number of protein–protein interactions of a protein correlates negatively with its rate of evolution. This relationship exists in spite of high levels of noise in interaction data sets [19] and it is not ablated when protein expression is accounted for by partial correlation.

Protein–protein-interaction databases are valuable in studies of protein evolutionary rate and in studies to determine how entire systems of interactions (protein modules) have evolved. By using the classification of yeast proteins into 'party' hubs and 'date' hubs [20], Fraser [21] showed that hubs that connect proteins of the same functional module evolve significantly slower than hubs that connect proteins from different functional modules. This study illustrates a fourth plausible hypothesis of evolutionary rate: that (at least some) proteins evolve according to the type of functional module they occupy. How the systems-level properties of protein interactions exert either constraining or creative evolutionary forces is an exciting direction for future investigation, and deserves further attention. However, it is likely that these higher-level studies will be more sensitive to covariates, which must be considered in appropriate experimental design.

A general model of protein evolution

It is clear from recent studies that each variable (dispensability, expression, functional constraint and modularity) has a separate, important role in protein evolution (Figure 1). The exact nature of the relationships between these variables remains unclear, and more work is needed before they converge in a general, comprehensive model of protein evolution. This lack of clarity occurs partly because most research to date has focused on yeast, for the simple reason that this is the eukaryotic organism for which we have the most functional genomic data. However, genomic databases from a single organism are unlikely to apply across a wide phylogenetic breadth. For example, only 61% of the genes that are essential in *S. cerevisiae* are essential in *Candida albicans* [22]. Zhang and He [13] have demonstrated that this problem of

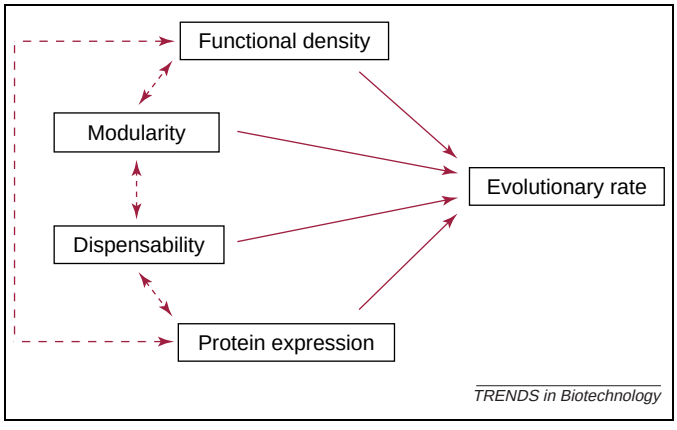

Figure 1. A model of protein evolution. Dashed arrows represent marginal correlation between the independent functional genomic variables (Table 1) and unbroken arrows represent the independent, significant effects of each variable on the rate of protein evolution. Without better analytical methods, we cannot be certain of the relative importance of any one variable. At present, our best approximation is that each plays a significant role in directing protein change over evolutionary time.

phylogenetic scale can weaken and, potentially, obscure conclusions about protein evolution if they are made using relatively divergent species. Thus, the ubiquity of this part of the model of protein evolution (Figure 1) should be treated cautiously until additional dispensability data sets are analyzed in more taxa.

Conclusions and future directions

The technological advances in the development of functional genomic databases, and improvements in analytical methods to measure comparative genomic variables such as codon bias and evolutionary rate [10,23] have begun to greatly improve our understanding of the mechanisms driving protein evolution. This improvement comes despite the often very low signal-to-noise ratio in some genomic variables, particularly measures of protein dispensability. Novel ways to account for noise, using either statistical modeling or cross-referencing genomic databases [10], will apply beyond the field of molecular evolution. Although in the near future it is possible that high-throughput genomics data will be measured with negligible error, it is encouraging that effective methods for handling that error exist currently.

A further consideration is that the variables presented here might have different levels of impact on subsets of proteins within the proteomes of organisms. For example, the rate of evolution of hubs might be predicted best by properties such as modularity, whereas the evolution of other proteins, such as loosely connected proteins in the interaction map, might be predicted best by either dispensability or expression. Dissecting the proteome into its constituent parts will shed light on this prospect.

Although we are closer to a better understanding of the mechanistic causes of protein evolutionary rate, we cannot be certain of this without examining the generality of these correlations in a wider sample of organisms. As pointed out previously, dispensability measurements from one lineage are unlikely to apply to taxa beyond immediate sister lineages. This problem of phylogenetic scale is likely to apply to similar studies that use large bioinformatic databases. Future research directed at sampling protein–protein interaction networks, dispensability measurements and other functional genomic variables from a taxonomically diverse set of genomes will help to determine the generality of the model that we describe here.

Acknowledgements

We thank Matthew Henn for helpful suggestions in the preparation of this article.

References

1 Kimura, M. and Ohta, T. (1974) On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* 71, 2848–2852
2 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
3 Fraser, H.B. *et al*. (2004) Coevolution of gene expression among interacting proteins. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9033–9038
4 Fraser, H. *et al*. (2003) A simple dependence between evolution rate and the number of protein–protein interactions. *BMC Evol. Biol.* 3, 11
5 Jordan, I.K. *et al*. (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
6 Pal, C. *et al*. (2003) Genomic function: rate of evolution and gene dispensability. *Nature* 421, 496–497
7 Yang, J. *et al*. (2003) Rate of protein evolution versus fitness effect of gene deletion. *Mol. Biol. Evol.* 20, 772–774
8 Akashi, H. (2003) Translational selection and yeast proteome evolution. *Genetics* 164, 1291–1303
9 Akashi, H. and Gojobori, T. (2002) Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc. Natl. Acad. Sci. U. S. A.* 99, 3695–3700
10 Wall, D.P. *et al*. (2005) Functional genomic analysis of the rates of protein evolution. *Proc. Natl. Acad. Sci. U. S. A.* 102, 5483–5488
11 Kellis, M. *et al*. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423, 241–254
12 Deutschbauer, A.M. *et al*. (2005) Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* 169, 1915–1925
13 Zhang, J. and He, X. (2005) Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol. Biol. Evol.* 22, 1147–1155
14 Wang, Y. *et al*. (2002) Precision and functional specificity in mRNA decay. *Proc. Natl. Acad. Sci. U. S. A.* 99, 5860–5865
15 Wilson, A.C. *et al*. (1977) Biochemical evolution. *Annu. Rev. Biochem.* 46, 573–639
16 Zuckerkandl, E. (1976) Evolutionary processes and evolutionary noise at the molecular level: I. Functional density of proteins. *J. Mol. Evol.* 7, 167–183
17 Vidalain, P.O. *et al*. (2004) Increasing specificity in high-throughput yeast two-hybrid experiments. *Methods* 32, 363–370
18 Fraser, H. *et al*. (2002) Evolutionary rate in the protein interaction network. *Science* 296, 750–752
19 von Mering, C. *et al*. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417, 399–403
20 Han, J.D. *et al*. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 430, 88–93
21 Fraser, H.B. (2005) Modularity and evolutionary constraint on proteins. *Nat. Genet.* 37, 351–352
22 Bruno, V.M. and Mitchell, A.P. (2004) Large-scale gene function analysis in *Candida albicans*. *Trends Microbiol.* 12, 157–161
23 Hirsh, A.E. *et al*. (2005) Adjusting for selection on synonymous sites in estimates of evolutionary distance. *Mol. Biol. Evol.* 22, 174–177
24 Hirsh, A.E. and Fraser, H.B. (2001) Protein dispensability and rate of evolution. *Nature* 411, 1046–1049
25 Hurst, L.D. and Smith, N.G. (1999) Do essential genes evolve slowly? *Curr. Biol.* 9, 747–750
26 Jordan, I.K. *et al*. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* 12, 962–968
27 Pal, C. *et al*. (2001) Highly expressed genes in yeast evolve slowly. *Genetics* 158, 927–931
28 d'Enfert, C. *et al*. (2005) CandidaDB: a genome database for *Candida albicans* pathogenomics. *Nucleic Acids Res.* 33, D353–D357
29 Giaever, G. *et al*. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418, 387–391
30 Ito, T. *et al*. (2002) Roles for the two-hybrid system in exploration of the yeast protein interactome. *Mol. Cell. Proteomics* 1, 561–566
31 Schwikowski, B. *et al*. (2000) A network of protein-protein interactions in yeast. *Nat. Biotechnol.* 18, 1257–1261
32 Salwinski, L. *et al*. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* 32, D449–D451
33 Uetz, P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature* 403, 623–627
34 Barrett, T. *et al*. (2005) NCBI GEO: mining millions of expression profiles database and tools. *Nucleic Acids Res.* 33, D562–D566
35 Holstege, F.C. *et al*. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95, 717–728
36 Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556

0167-7799/$ - see front matter
doi:10.1016/j.tibtech.2005.07.009