EVOLUTION AFTER GENE DUPLICATION

Edited by

Katharina Dittmar
SUNY at Buffalo
Buffalo, New York

David Liberles
University of Wyoming
Laramie, Wyoming

WILEY-BLACKWELL

A JOHN WILEY & SONS, INC., PUBLICATION

11 On the Energy and Material Cost of Gene Duplication

ANDREAS WAGNER

Department of Biochemistry, University of Zurich, Zurich, Switzerland; The Santa Fe Institute, Santa Fe, New Mexico; Swiss Institute of Bioinformatics, Lausanne, Switzerland

1 INTRODUCTION

A gene duplication first occurs in a single individual of an evolving population. The duplicate may then increase in frequency or again become extinct. Genetic drift and natural selection may be responsible for either fate. If natural selection is involved, one must distinguish two principal contributors to this fate: duplication benefits and duplication costs.

Gene duplication has long- and short-term evolutionary benefits. Among the long-term benefits is the ability to facilitate evolutionary innovation through the evolution of new molecular activities in one of the gene copies, a notion first popularized by Ohno (1970). However, such long-term benefits may be irrelevant for the immediate fate of a gene duplicate after it first arises. Shorter-term benefits include advantages of increased gene dosage and thus increased gene expression. Such advantages may exist both for gene products that are in extremely high demand in a cell, and for genes that are expressed at very low levels when in single copy. In the latter case, noisy gene expression is at the root of the benefit. Noisy gene expression is ubiqituous, but especially prevalent for lowly expressed genes (Bar-Even et al., 2006). For such genes, the amount of gene product in a cell can show dramatic fluctuations, and for long periods of time the cell may contain little or none of the product. If the product is important to the life cycle of a cell, it is advantageous to alleviate these fluctuations via an increase in the average expression level (Cook et al., 1998). Gene duplication is one avenue to such an increase. Another short-term benefit arises in cases where a gene's duplicate is not equal in sequence and function to the original. If the new function is beneficial to the cell, its carrier may rise in frequency through natural selection. Both anecdotal evidence (Long and Langley, 1993) and systematic work on genome-scale data (Katju and Lynch, 2003; Vinckenbosch et al., 2006) show that new genes can indeed originate in this way.

The second factor influencing a gene duplication's fate through natural selection is the cost of a duplication. A duplication will generally result in an increase in a

Evolution After Gene Duplication, Edited by Katharina Dittmar and David Liberles

cell's genome size. This may result in an increased amount of time needed for DNA replication (and cell division), as well as in additional energy and material needs for DNA replication. As a result, cells with only a single copy of any one gene might be able to divide slightly faster. This cost component, however, is likely to play only a minor role. The generally small increase in genomic DNA associated with a single-gene duplication might cause a small replication delay in prokaryotes with a single replication origin, but not so in eukaryotes, where DNA replication is initiated simultaneously at thousands of replication origins in the genome. For example, the genome of *Xenopus laevis* is approximately 1000 times larger than that of that in *Escherichia coli*. Nonetheless, it can replicate in some 30 minutes, not much longer than the minimum cell division time of *E. coli* (Alberts, 2002). In addition, the energy and material cost of synthesizing the added DNA is negligible compared to that of gene expression. For example, dividing yeast (*Saccharomyces cerevisiae*) cells can double their biomass every 90 minutes. Fifty percent of this biomass consists of protein and RNA, but only 0.4% consists of DNA (Forster et al., 2003).

Two other cost components are likely to be more important than a gene duplication's influence on genome size. Both stem from the increase in expression caused by a duplication. While cells may compensate for changes in gene dosage by adjusting expression levels (Kafri and Pilpel, 2004)—for example, through negative feedback regulation of the duplicated gene, or via limited availability of transcription factors—such mechanisms may not be prevalent (Wong and Roth, 2005; He and Zhang, 2006). In the absence of such mechanisms, one would expect an approximate doubling of a gene's expression level after duplication, if a regulatory region is duplicated in its entirety along with the coding region. Increased gene expression may interfere with cellular life in a variety of ways. For example, the newly expressed gene product may bind to proteins that are then no longer available for other, necessary protein interactions. This is one of several ways in which increased gene expression may be *toxic* to a cell. Second, gene expression requires both energy (in the form of ATP) and materials (nucleotides and amino acids) which incur a cost on a cell's energy budget or material budget, whenever this budget is limited.

It is very difficult to disentangle the relative contributions of expression toxicity and its energy or material cost, partly because toxicity has many faces. I will discuss recent evidence from the yeast *S. cerevisiae* that gene expression cost alone—even disregarding potentially toxic effects of increased expression—can affect the fate of most duplicates, at least in organisms with large population sizes. Before that, however, I need to ask how small an expression cost can be visible to natural selection.

2 COSTS VISIBLE TO NATURAL SELECTION

The fitness cost of any mutation, including gene duplications, is typically expressed in terms of a selection coefficient s, a fitness reduction relative to the wild type that its carrier suffers. In a diploid organism, the magnitude of s below which genetic drift influences a mutation's fate more strongly than natural selection is $s < \frac{1}{4N_e}$ (Kimura, 1983). Here N_e is the effective size of a population, which can be estimated from the nucleotide diversity at synonymous sites. Existing nucleotide diversity data show that for yeast, the critical s below which drift is stronger than selection is smaller than 5×10^{-7} (Wagner, 2005, 2007; Bragg and Wagner 2007, 2009). This means that

minute effects of mutations, many orders of magnitude smaller than could be detected in the laboratory, can affect the fate of a mutation.

3 ENERGY COST OF GENE EXPRESSION IN THE YEAST *Saccharomyces cerevisiae*

A dividing cell needs a certain amount of energy per division cycle, much of it invested in building cell biomass. It is reasonable to assume that the production of such energy is one of the limiting factors in cell proliferation. If so, then increasing the expression of any one gene leaves less of this energy for growing the remaining biomass, which would delay cell proliferation by an amount corresponding to the fraction of energy diverted to the gene's expression. Thus, the fractional energy cost of expressing any one gene is an indicator of the fitness effect s that a duplication of this gene has, in situations where cell growth rate is proportional to fitness. I note that gene expression itself is responsible for a substantial fraction of biomass production. As mentioned above, in yeast, RNA and protein comprise fully half of a cell's biomass (Forster et al., 2003).

The energy cost of gene expression has many components. First, nucleotide precursors need to be synthesized, which carries a cost in terms of both material and energy. Second, these nucleotide precursors need to be strung together in transcription to make messenger RNA. Third, amino acids need to be synthesized. Fourth, these amino acids need to be polymerized in translation. In addition, one needs to take into account different rates of protein and RNA turnover. Both kinds of molecules are constantly synthesized and degraded at molecule-specific rates that vary over several orders of magnitude (Wang et al., 2002; Belle et al., 2006). The absolute steady-state concentration of an RNA and protein molecule is thus not very informative about its expression cost. A molecule might experience fast synthesis and high decay rates, or slow synthesis and low decay rates, both of which might yield the same steady state, but at very different cost to a cell. In sum, to estimate the expression cost of genes, we need information about precursor synthesis costs, synthesis rates, and half-life. Currently, such information is available on a genome scale for only one organism, the yeast *S. cerevisiae*.

By integrating a vast amount of genome-scale information on mRNA and protein levels, mRNA and protein half-lives, nucleotide composition of genes, and nucleotide and amino acid synthesis costs, one can determine what fraction of a cell's gene expression energy cost goes into the expression of any one gene. The result is a distribution of selection coefficients associated with doubling the expression for each of thousands of yeast genes (Figure 1) (Wagner, 2005, 2007). Strikingly, all yeast genes for which expression information is available have expression costs vastly greater than the critical s discussed above. This holds regardless of whether the cells grow under fermentative or respiratory conditions. This means that for yeast genes expressed at any level, duplication would generally carry a cost visible to selection. To be sure, this assertion relies on some assumptions, among them that the energy cost of producing RNA and protein biomass is not vastly different from that of the cell's remaining biomass (among it, many lipids and sugars). However, even if all selection coefficients in Figure 1 were overestimated tenfold, duplication of most yeast genes would still be subject to costs visible to selection.

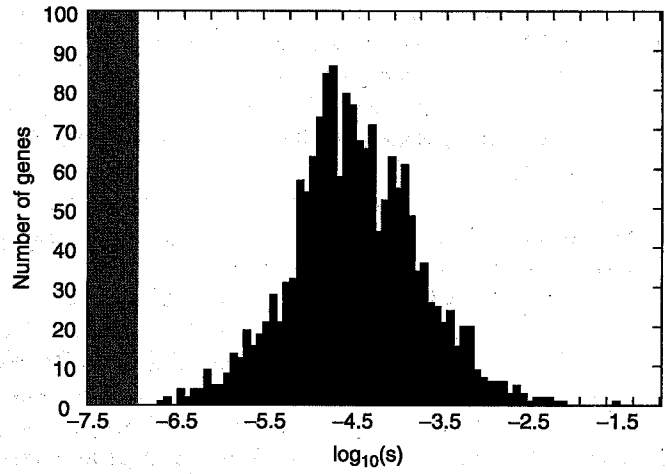

Figure 1 Distribution of the fractional energy cost s of doubling gene expression for the yeast *S. cerevisiae*. The gray zone indicates a region where the cost is too small to be visible to natural selection, based on effective population size estimates of yeast. (After Wagner, 2007.)

4 MATERIAL COST OF GENE EXPRESSION IN THE YEAST *SACCHAROMYCES CEREVISIAE*

Some elements are major components of the biomass produced in gene expression. Specifically, RNA contains carbon, nitrogen, and phosphorus. Protein contains carbon, nitrogen, and sulfur. These elements can severely restrict the growth of organisms when their availability is limited. Such limitation can also foster fierce competition. In an environment where any one element is limiting, an increase in expression of any one gene will divert elemental nutrients to the gene product and may thus reduce the rate of cell proliferation. Because the chemical compositions of amino acids and nucleotides are known, and because we have complete genome sequence information, we can determine the amount of any one element invested into a single RNA or protein molecule. In combination with the known biomass composition of yeast, and with available information on mRNA and protein expression levels and half-lives, we can thus determine, for each element and gene, the material cost of doubling gene expression. This cost can be expressed as a fraction s of a cell's estimated total material budget. By relating s to a critical selection coefficient, as outlined above, one can determine whether a given cost increase is visible to natural selection (Bragg and Wagner, 2007, 2009). With this approach, one finds that for more than 97% of yeast genes and for the elements carbon, nitrogen, and sulfur, the cost of doubling expression is a factor of 10 greater than the critical selection coefficient. The effect of phosphorus limitation is less dramatic, being visible for only 94% of duplicated genes. These numbers change if any one element is not strongly but weakly limiting. For example, if a fractional increase in expression cost by x causes a reduction in fitness not by x but merely by $x/4$, a doubling of expression would be visible to selection only for more than 90% of genes. In sum, for any element that is growth limiting, gene duplication causes significant material costs for the vast majority of genes, similar to what I discussed earlier for energy costs.

Energy cost and material cost of a gene's expression are highly positively correlated (Bragg and Wagner, 2007, 2009). Genes with a high energy cost of expression also tend to have a high material cost. It is easy to see why. A substantial part of both costs comes from the rate of synthesis for mRNA and protein molecules, which enter the calculation of both energy and material in identical ways. An additional contribution to this correlation comes from the fact that chemically complex amino acids, containing more atoms of a given type, tend to consume more energy in biosynthesis than simpler amino acids. (The cost differences among different nucleotides are much smaller than those among different amino acids and are thus less important.)

5 THE *LAC* OPERON AS AN EXPERIMENTAL SYSTEM TO STUDY EXPRESSION COSTS

I now highlight some recent experimental work on the *lac* operon that sheds light on the cost of expression for very highly expressed genes. The *lac* operon is one of the best-studied regulatory systems inside cells (Alberts, 2002). Its three gene products are a β-galactosidase (product of the *lacZ* gene), a permease (*lacY*), and a transacetylase (*lacA*). The first two of these products are necessary to metabolize the sugar lactose. The expression of the *lac* operon is highly regulated and turned on only if lactose is available in the cell's environment. In such environments, the operon is expressed at very high levels. The advantage of this system is that its regulation can be manipulated either through mutations or through artificial inducers. One such inducer is isopropyl-β-D-thiogalactoside (IPTG). IPTG induces the *lac* operon, but the cell does not gain any benefit from this induction, because unlike lactose, IPTG cannot feed into energy metabolism. A recent study (Dekel and Alon, 2005) took advantage of this property to measure the cost of expressing the *lac* operon at various levels of induction. It concluded that full induction of the *lac* operon with IPTG leads to a reduction in the cell division rate of 4.5%. Although this type of approach cannot strictly exclude the possibility that the cost of expression reflects toxicity of the gene products, this seems unlikely in the case of the *lac* operon. The reason is that the high expression state is not just induced in the laboratory under unphysiological conditions with an artificial inducer, but it is also vital under physiological conditions in lactose-containing environments.

Another study took advantage of mutations that render *lacZ* expression constitutive (Stoebel et al., 2008). It is estimated that *lac* operon expression in lactose-free environments leads to a 10% reduction in growth rate. Most of this cost comes from expressing β-galactosidase (Stoebel et al., 2008). Tagging the β-galactosidase product with a peptide that decreases its half-life and thus recycles its amino acids reduces this cost dramatically. This suggests that the bulk of the cost for expressing this protein does not come from the biosynthesis of the proteins and its amino acids. Aside from the possibility that the cost of transcription is of major importance, it is also conceivable that the extremely high *lac* expression sequesters RNA polymerases or ribosomes, rendering them unavailable for expressing other genes at appropriate levels.

Experimental approaches like these are powerful, because they can demonstrate the effects of gene expression on cell growth directly. However, they can detect the expression costs of only the most highly expressed genes, because experiments are able to resolve selection coefficients only to a lower limit of approximately 10^{-3}.

In organisms with large effective population size, much smaller selection coefficients are still visible to selection. Importantly, most genes have small selection coefficients associated with a doubling of gene expression. In yeast, doubling the expression of most genes would lead to expression costs much smaller than 10^{-3}. The example just discussed also shows that for the enormous changes in expression that occur in the *lac* operon, factors independent of material or energy cost, such as the sequestering of polymerases or ribosomes, may come into play. These factors may play a smaller role for more lowly expressed genes and for smaller expression changes, such as those observed in a gene's duplication.

6 EVOLUTIONARY COST SIGNATURES

Where experiments cannot reach, patterns of evolutionary change may inform us about the impact of duplication costs. A genome-scale analysis of gene duplicates in yeast shows that genes with high carbon and nitrogen expression cost have fewer surviving duplicates (Bragg and Wagner, 2007). In such an analysis, it is important to correct for gene expression levels, because genes with high expression may also evolve a nucleotide composition with low elemental or energy cost (Akashi and Gojobori, 2002; Fauchon et al., 2002; Elser et al., 2006; Heizer et al., 2006). However, the association persists when differences in expression levels are taken into account (Bragg and Wagner, 2007). In addition to this example pertaining to gene duplications, a number of studies have demonstrated that energetic and material costs of expression shape the composition of proteins. For example, Akashi and Gojobori (2002) showed that in *E. coli* highly expressed proteins show increased abundance of energetically cheap amino acids. In addition, proteins needed to assimilate carbon tend to contain fewer carbon-costly amino acids than other proteins (Baudouin-Cornu et al., 2001). A similar pattern holds for proteins involved in sulfur assimilation (Baudouin-Cornu et al., 2001). These patterns probably reflect an evolutionary adaptation which ensures that nutrient assimilation can remain active if a nutrient becomes scarce.

As discussed earlier, expression cost is only one of multiple factors affecting the fate of duplicate genes. That it can leave genomic signatures at all is thus astounding. It suggests that expression cost has a strong influence on molecular evolution. Benefits of duplication, however, can also leave genomic signatures. For example, highly active metabolic enzymes (i.e., metabolic enzymes with high metabolic flux) tend to be encoded by a greater number of duplicate genes than are less active enzymes (Papp et al., 2004; Vitkup et al., 2006). This pattern probably reflects the advantage of increased gene dosage for such enzymes, an advantage that may override their large expression cost. The types of signatures gene duplication leaves in a genome reflect whether a duplicate's fate is dominated by either benefit or cost.

7 CONCLUSIONS

In microbial organisms, the doubling of expression associated with many gene duplications carries significant energetic and material costs. Such duplications thus do not go to fixation neutrally. Because most genomes contain large numbers of duplicate genes, one can infer that gene duplication often confers adaptive advantages that outweigh

these costs. To investigate the nature of these advantages is one part of a promising research program that will yield insight into the evolutionary forces shaping genomes. Another part is the investigation of expression costs in higher, multicellular organisms. Because of their smaller effective population sizes, selection is a weaker evolutionary force in these organisms. It is currently unclear whether the observations discussed here apply to higher organisms.

Acknowledgments

I thank the Swiss National Foundation for support through SNF grant 315200-116814 and through the YeastX program from SystemsX.ch.

REFERENCES

Akashi H, Gojobori T. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc Natl Acad Sci USA 99:3695–3700.

Alberts B. 2002. *Molecular Biology of the Cell*. New York: Garland Science.

Bar-Even A, Paulsson J, et al. 2006. Noise in protein expression scales with natural protein abundance. Nat Genet 38:636–643.

Baudouin-Cornu P, Surdin-Kerjan Y, et al. 2001. Molecular evolution of protein atomic composition. Science 293:297–300.

Belle A, Tanay A, et al. 2006. Quantification of protein half-lives in the budding yeast proteome. Proc Natl Acad Sci USA 103:13004–13009.

Bragg J, Wagner A. 2007. Protein carbon content evolves in response to carbon availability and may influence the fate of duplicate genes. Proc R Soc Lond Ser B 274:1063–1070.

Bragg J, Wagner A. 2009. Protein material costs: single atoms can make an evolutionary difference. Trends Genet 25:5–8.

Cook DL, Gerber LN, et al. 1998. Modeling stochastic gene expression: implications for haploinsufficiency. Proc Natl Acad Sci USA 95:15641–15646.

Dekel E, Alon U. 2005. Optimality and evolutionary tuning of the expression level of a protein. Nature 436:588–592.

Elser JJ, Fagan WF, et al. 2006. Signatures of ecological resource availability in the animal and plant proteomes. Mol Biol Evol 23:1946–1951.

Fauchon M, Lagniel G, et al. 2002. Sulfur sparing in the yeast proteome in response to sulfur demand. Mol Cell 9:713–723.

Forster J, Famili I, et al. 2003. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. Genome Res 13:244–253.

He XL, Zhang JZ. 2006. Transcriptional reprogramming and backup between duplicate genes: Is it a genomewide phenomenon? Genetics 172:1363–1367.

Heizer EM, Raiford DW, et al. 2006. Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis. Mol Biol Evol 23:1670–1680.

Kafri RB-E, Pilpel Y. 2004. Transcription control reprogramming in genetic backup circuits. Nat Genet 37:295–299.

Katju V, Lynch M. 2003. The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. Genetics 165:1793–1803.

Kimura M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge; UK: Cambridge University Press.

Long MY, Langley CH. 1993. Natural-selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. Science 260:91–95.

Ohno S. 1970. *Evolution by Gene Duplication*. New York: Springer-Verlag.

Papp B, Pál C, et al. 2004. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. Nature 429:661–664.

Stoebel D, Dean A, et al. 2008. The cost of expression of *Escherichia coli lac* operon proteins is in the process, not the products. Genetics 178:1653–1660.

Vinckenbosch N, Dupanloup I, et al. 2006. Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci USA 103:3220–3225.

Vitkup D, Kharchenko P, et al. 2006. Influence of metabolic network structure and function on enzyme evolution. Genome Biol 7:R39.

Wagner A. 2005. Energy constraints on the evolution of gene expression. Mol Biol Evol 22:1365–1374.

Wagner A. 2007. Energy costs constrain the evolution of gene expression. J Exp Zool B 308B:322–324.

Wang YL, Liu CL, et al. 2002. Precision and functional specificity in mRNA decay. Proc Natl Acad Sci USA 99:5860–5865.

Wong SL, Roth FP. 2005. Transcriptional compensation for gene loss plays a minor role in maintaining genetic robustness in *Saccharomyces cerevisiae*. Genetics 171:829–833.