Letter to the Editor

Inferring Lifestyle from Gene Expression Patterns

Andreas Wagner

Department of Biology, University of New Mexico and Santa Fe Institute

For many organisms, the primary locus of study is the laboratory, and not the organism's natural habitat. Through a century of laboratory studies, a huge body of knowledge has been accumulated for several ''model organisms'' of molecular and cell biology. This contrasts sharply with the often limited amount of information available on the ecology of such model organisms, a discrepancy that is particularly striking for microbes. Microbes arguably provide the bulk of our cell biological knowledge, but their natural habitats are poorly understood. Their physiology, their genomic gene content, and the structure of their genetic networks have been shaped over millions of years by natural selection in the wild. However, even for model microbes such as *Escherichia coli* and yeast, little is known about the ecological conditions under which they evolved. And because of the difference in laboratory and natural environments, laboratory experiments often have limited value in providing an understanding of these conditions. A case in point is the huge number of gene knockout experiments in multiple eukaryotes that show little or no phenotypic effects in the laboratory (Tautz 1992; Smith et al. 1997; Wagner 2000). The artificial conditions under which these experiments are carried out may sometimes be responsible for the absence of such phenotypic defects. Arguably, in the wild, such knockout mutations might be eliminated from the population. It would thus be best to assay their effects under more realistic conditions.

Even when known, the complexity of a natural environment, such as the vertebrate intestine for *E. coli*, is difficult to emulate in the laboratory. However, experimenters often have a broad range of choices of laboratory conditions under which to study an organism. Some of them may more closely resemble the situation in the wild and should thus be preferred over others. Especially for microbes, a number of simple and fundamental choices are possible. What is the main carbon source, if any, available to the organism in the wild? Is energy metabolism predominantly aerobic or anaerobic? Is there a dominant nitrogen source? For organisms that are facultatively diploid, in which ploidy stage do they spend the majority of their life cycle? Are they frequently or rarely exposed to DNA-damaging agents such as UV light? Answers to such questions have implications far beyond the choice of a suitable laboratory environment. They provide fundamental insights into an organism's ecology. This paper offers a suggestion as to how these questions might be addressed without extensive ecological studies, through an approach whose key ingredient is information on codon usage bias in completely sequenced genomes.

Codon usage bias is the preferential occurrence of particular codons for amino acids that are encoded by more than one codon. In microbes, preferred codons are those for which the respective tRNAs are abundant. Highly expressed genes have highly biased codon usage, which ensures efficient translation. Genes expressed at a lower level tend to show less selective codon occurrence. Because the expression level of each gene depends on the environment, the observed distribution of codon usage bias should reflect gene expression levels in a typical environment or the mix of environments encountered by the organism on an evolutionary timescale. Gene expression levels of many genes and their codon biases would be highly correlated in a (laboratory) environment or a mix of environments similar to that in which the organism evolved. Conversely, in an environment that is very dissimilar to that typically encountered by an organism, the correlation between codon usage bias and gene expression levels will be poor.

Large-scale gene expression studies have shown that even seemingly simple physiological changes entail expression changes in vast numbers of genes. A case in point is the diauxic shift in the yeast *Saccharomyces cerevisiae*, which is the change from anaerobic (fermentative) to aerobic (respiratory) metabolism as a cell depletes its fermentable carbon source (such as glucose) and has to rely on a nonfermentable carbon source such as ethanol. During the diauxic shift, the mRNA expression level of more than 1,700, or 27% of all yeast genes, changes by more than a factor of two (DeRisi, Iyer, and Brown 1997). One of the most basic questions one can ask about the life of a facultatively anaerobic organism such as yeast is whether its metabolism in the wild is predominantly fermentative or oxidative. For those organisms that show a strong Pasteur effect (the inhibition of fermentation by oxygen), this also suggests a question about the abundance of oxygen in the environment in which they evolved.

Figure 1 shows a scatterplot of yeast gene expression level versus codon usage bias, as measured by the codon bias index (CBI; Bennetzen and Hall 1981). Shown are mRNA expression levels under fermentative (fig. 1*a*) and oxidative (fig. 1*b*) conditions of nonribosomal yeast genes that show both a significant codon usage bias and a significant change in expression during the diauxic shift. Codon usage bias is significantly correlated with gene expression levels only for cells growing fermentatively. Thus, based on this assay, the codon bias indices observed in yeast have probably evolved largely under the influence of fermentation, and fermentable carbon sources may thus be yeast's prevalent

Abbreviation: CBI, codon bias index.

Key words: microarray, microbial ecology, genomics.

Address for correspondence and reprints: Andreas Wagner, Department of Biology, University of New Mexico, 167A Castetter Hall, Albuquerque, New Mexico 817131-1091. E-mail: wagnera@unm.edu.

Mol. Biol. Evol. 17(12):1985–1987. 2000

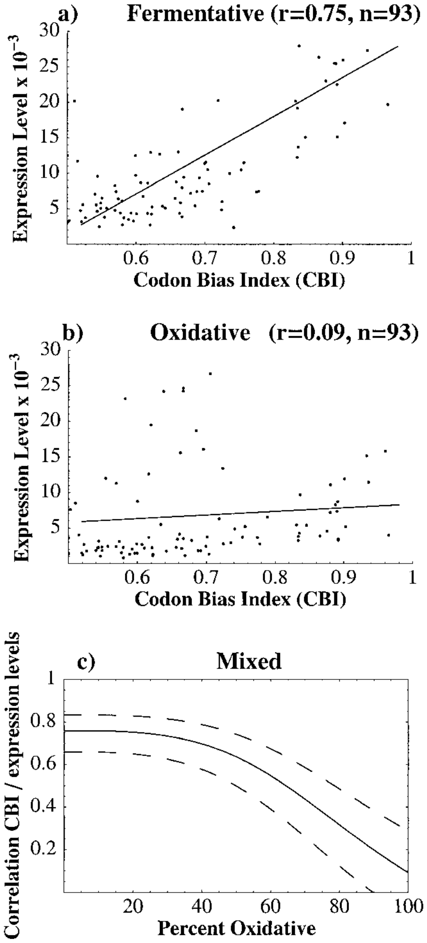

FIG. 1.—Codon bias index (CBI; Bennetzen and Hall 1981) versus expression level for 93 yeast genes that show a greater than twofold change in mRNA expression level during the shift from anaerobic to aerobic metabolism. *a*, Expression level versus CBI before the diauxic shift (Pearson $r = 0.75$; $P < 10^{-6}$; Kendall $\tau = 0.50$) *b*, Expression level versus CBI after the diauxic shift (Pearson $r = 0.09$; $P > 0.2$; Kendall $\tau = 0.27$). A *t*-test after Fisher's *z*-transformation of the Pearson *r* values indicates that the two correlation coefficients are significantly different at $P < 10^{-4}$. Expression values shown represent absolute fluorescence intensity (minus background) at time step 7 of the diauxic shift experiment reported by DeRisi, Iyer, and Brown (1997), where fermentative and oxidative expression levels are taken from the Cy3-dUTP-labeled control population and the Cy5-dUTP-labeled population, respectively. Because of the considerable variation across microarray experiments, expression levels shown do not translate into absolute mRNA concentrations and are only informative when viewed in relation to other genes. Codon bias indices range from −1 to 1, where a CBI of 0 indicates no codon bias and a CBI of 1 indicates the most severe codon bias associated with the most highly expressed genes. Negative codon bias levels are rarely observed. Only nonribosomal yeast genes with significant codon bias (CBI > 0.5) are included in the analysis, but qualitative results are similar if all genes are included. Ribosomal proteins are excluded here because they are highly expressed under many conditions. For the purpose of comparison, $r^2 = 0.14$ for all yeast genes whose expression levels do not change by a factor greater than two. *c*, For each gene whose expression level is shown in *a* and *b*, the expression levels before and after the diauxic shift (x_{bef} and x_{aft}) were used to calculate a linear interpolation, $y(t)$, between these values that estimates the average expression level of the gene if a cell spends *t* percent of its time in an oxidative state and ($100 - t$) percent in a fermentative state ($y(t) = (1 - t/100) \times x_{\mathrm{bef}} + (t/100) \times x_{\mathrm{aft}}$). For each value of *t* between 0 and 100, the Pearson correlation coefficient between the (interpolated) expression level and the CBI was calculated and is plotted as a function of *t* in the figure.

←

Dashed lines indicate 95% confidence intervals. The figure demonstrates that the assumption that the cell spends only a fraction of time in either state does not improve the correlation between CBI and expression level.

carbon sources. However, because yeast cells in the wild certainly cycle between metabolic states, the relationship between gene expression levels and codon usage bias may reflect this mix of states. This issue is addressed by figure 1*c*, which shows the correlation between CBI and expression level if a cell spends, on average, *t* percent of its time in an oxidative state. The figure is based on a numerical interpolation of gene expression levels between the two pure states shown in figure 1*a* and *b*, and it shows that no mixed state improves the correlation between expression level and CBI.

There are, of course, caveats to this approach. Methodologically, a major concern is the noisiness of microarray expression data and the limited correlation between mRNA and protein expression levels. However, the amount of noise is smaller and the correlation higher for highly expressed genes (Gygi et al. 1999), which are of most interest here. Second, aspects of fitness that do not leave a CBI signature will elude this approach. These aspects may include fitness components determined by an organism's interaction with its biotic environment (predation, competition, etc.) as opposed to its abiotic environment. Similarly, the importance of particular physiological states, such as quiescence, where gene expression is much reduced cannot be assessed. While many microbes in the wild may spend substantial amounts of time in quiescence (Lewis and Gattie 1991), it is also clear that periods of high cell activity must be evolutionarily important. Otherwise, we would not see extreme codon bias in highly expressed microbial genes. Third, it is possible that some parameters influencing translational efficiency, such as the distribution of tRNA species, change in different environments. Fourth, in highly derived laboratory strains of organisms, some evolution may have occurred in the laboratory (Ferea et al. 1999). However, for many poorly studied microbes, such as the increasing number of completely sequenced extremophiles, this approach may provide valuable information on their ecology that cannot be obtained otherwise.

LITERATURE CITED

BENNETZEN, J. L., and B. D. HALL. 1981. Codon selection in yeast. J. Biol. Chem. **257**:3026–3031.

DERISI, J. L., V. R. IYER, and P. O. BROWN. 1997. Exploring the metabolic and genetic-control of gene-expression on a genomic scale. Science **278**:680–686.

FEREA, T. L., D. BOTSTEIN, P. O. BROWN, and R. F. ROSENZWEIG. 1999. Systematic changes in gene expression patterns following adaptive evolution in yeast. Proc. Natl. Acad. Sci. USA **96**:9721–9726.

GYGI, S. P., Y. ROCHON, B. R. FRANZA, and R. AEBERSOLD. 1999. Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. **19**:1720–1730.

LEWIS, D. L., and D. K. GATTIE. 1991. The ecology of quiescent microbes. ASM News **57**:27–32.

SMITH, V., K. N. CHOU, D. LASHKARI, D. BOTSTEIN, and P. O. BROWN. 1997. Functional analysis of the genes of yeast chromosome V by genetic footprinting. Science **275**:464–464.

TAUTZ, D. 1992. Redundancies, development and the flow of information. Bioessays **14**:263–266.

WAGNER, A. 2000. The role of pleiotropy, population size fluctuations, and fitness effects of mutations in the evolution of redundant gene functions. Genetics **154**:1389–1401.

SIMON EASTEAL, reviewing editor

Accepted August 23, 2000