

Inferring life style from gene expression patterns

Andreas Wagner

University of New Mexico

and

The Santa Fe Institute

University of New Mexico

Department of Biology

167A Castetter Hall

Albuquerque, NM 817131-1091

Phone: +505-277-2021

FAX: +505-277-0304

wagnera@unm.edu

For many organisms, and especially for the model organisms of molecular and cell biology, the primary locus of study is the laboratory, and not their natural habitat. Thus, a huge body of knowledge accumulated through a century of laboratory studies contrasts with the limited amount of information available on the ecology of many model organisms. This discrepancy is particularly striking for microbes, which arguably provide the bulk of our cell biological knowledge, but whose natural habitats are poorly understood. Their physiology, their genomic gene content, and the structure of their genetic networks has been shaped over millions of years by natural selection in the wild. Because of this discrepancy, certain laboratory experiments have limited value in understanding the workings of an organism. A case in point are the huge number of gene knock-out experiments in multiple eukaryotes that show little or no phenotypic effects in the laboratory¹⁻³. 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 knock-out mutations might be eliminated from the population. It would thus be best to assay their effects under more realistic conditions.

Even where known, the complexity of a natural environment, such as the vertebrate intestine for *E. coli*, is difficult to emulate in the laboratory. However, experimentors often have a broad range of choices of laboratory conditions under which to study an organism. Some of them may resemble more closely 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, if any, is the main carbon source 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 do have implications far beyond the choice of a suitable laboratory environment. They provide fundamental insights into an organisms ecology. The following is a suggestion of how these questions might be answered without extensive ecological studies, 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, *we can expect that the observed distribution of codon usage bias reflects gene expression levels in a typical environment or the mix of environments encountered by the organism on an evolutionary time scale.* 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 non-fermentable 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 two⁴. One of the most basic questions one can ask about the life of a facultatively anaerobic organism such as yeast is whether it lives predominantly under aerobic or anaerobic conditions in the wild. Fig. 1 shows a scatterplot of gene expression level vs. codon usage bias, as measured by the codon bias index⁵ (CBI). Shown are mRNA expression levels under anaerobic (Fig. 1a) and aerobic (Fig. 1b) conditions of non-ribosomal 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 anaerobically on glucose. Thus, based on this assay, the physiological state of greater evolutionary relevance is anaerobic growth. However, because yeast cells in the wild certainly cycle between anaerobic and aerobic states, the relation between gene expression levels and codon usage bias may reflect this mix of states. This issue is addressed by Fig. 1c, which shows the correlation between CBI and expression level if a cell spends, on average, t percent of its time in an aerobic state. The figure is based on a numerical interpolation of gene expression levels between the two pure states shown in Fig. 1a and 1b, and it shows that no mixed state improves the correlation between expression level and CBI.

Caveats to this approach include the noisiness of microarray expression data, and the limited correlation between mRNA and protein expression levels. Second, the approach may not be suitable to determine the importance of particular physiological states, such as stationary phase, where gene expression is much reduced. Third, it is possible that some parameters influencing translational efficiency, such as the distribution of tRNA species change in different environments. Also, when applied to highly derived laboratory strains, it is possible that one measures to some extent

evolution that has occurred in the laboratory. Thus, the approach should ideally be applied to specimens sampled from their natural habitat.

1. Tautz, D. Redundancies, Development and the Flow of Information. *Bioessays* **14**, 263-266 (1992).
2. Wagner, A. Redundant gene functions and natural selection. *Journal of evolutionary biology* **12**, 1-16 (1999).
3. Smith, V., Chou, K.N., Lashkari, D., Botstein, D. & Brown, P.O. Functional-Analysis of the Genes of Yeast Chromosome-V By Genetic Footprinting. *Science* **275**, 464-464 (1997).
4. Derisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the Metabolic and Genetic-Control of Gene-Expression On a Genomic Scale. *Science* **278**, 680-686 (1997).
5. Bennetzen, J.L. & Hall, B.D. Codon selection in yeast. *The Journal of Biological Chemistry* **257**, 3026-3031 (1981).

Fig. 1. Codon bias index ⁵ (CBI) vs. expression level of 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 vs. CBI before the diauxic shift (Pearson $r=0.75$; $P<10^{-6}$; Kendall $\tau=0.50$) **b)** Expression level vs. 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 seven of the diauxic shift experiment reported by ⁴, where anaerobic and aerobic expression levels are taken from the Cy3-dUTP labeled control population and the Cy5-dUTP labeled population, respectively. Because of the considerable variation across micro array 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 zero 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 non-ribosomal yeast genes with a 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.

c) For each gene whose expression level is shown in panel a) and b), the expression level before and after the diauxic shift (x_{bef} and x_{aft}), was used to calculate a linear interpolation, $y(t)$, between these values that estimates the average expression level of the gene if a cell spends $t\%$ of its time in an aerobic state, and $(100-t)\%$ in an anaerobic state ($y(t)=(1-t/100) \times x_{bef} + (t/100) \times x_{aft}$). For each value of t between 0 and 100, the Pearson correlation coefficient between the (interpolated) expression level and the CBI was calculated. It 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.