

# An Analysis of Determinants of Amino Acids Substitution Rates in Bacterial Proteins

Eduardo P. C. Rocha\*† and Antoine Danchin\*‡

\*Unité GGB, Institut Pasteur, Paris, France; †Atelier de BioInformatique, Université Pierre et Marie Curie, Paris, France; and ‡HKU-Pasteur Research Centre, Hong Kong

The variation of amino acid substitution rates in proteins depends on several variables. Among these, the protein's expression level, functional category, essentiality, or metabolic costs of its amino acid residues may play an important role. However, the relative importance of each variable has not yet been evaluated in comparative analyses. To this aim, we made regression analyses combining data available on these variables and on evolutionary rates, in two well-documented model bacteria, *Escherichia coli* and *Bacillus subtilis*. In both bacteria, the level of expression of the protein in the cell was by far the most important driving force constraining the amino acids substitution rate. Subsequent inclusion in the analysis of the other variables added little further information. Furthermore, when the rates of synonymous substitutions were included in the analysis of the *E. coli* data, only the variable expression levels remained statistically significant. The rate of nonsynonymous substitution was shown to correlate with expression levels independently of the rate of synonymous substitution. These results suggest an important direct influence of expression levels, or at least codon usage bias for translation optimization, on the rates of nonsynonymous substitutions in bacteria. They also indicate that when a control for this variable is included, essentiality plays no significant role in the rate of protein evolution in bacteria, as is the case in eukaryotes.

## Introduction

As species evolve, the rate of amino acid substitutions in proteins varies considerably among different protein families (Dickerson 1971). Such variability is common to all living species and reflects the different intensities and types of selection pressure acting on protein synthesis, localization, and function (Nei 2000). Several general factors have been proposed to play an important role in the rate of protein evolution. Among these, we explored the role of four selective constraints expected to drive the evolution of bacterial proteins: expression level, functional category, essentiality, and metabolic cost of amino acid residues.

Because high expression levels bias codon usage toward the use of optimal codons under exponential growth conditions, the calculation of some measure of the codon usage bias in proteins has been extensively used to account for protein expression levels in bacteria. Using such measures, recently coupled to transcriptome data analysis, higher expression levels were found to correlate with lower rates of protein evolution in *Escherichia coli* (Sharp 1991) and *Saccharomyces cerevisiae* (Pal, Papp, and Hurst 2001). The functional category of a protein may also constrain its rate of evolution, if only because different categories imply different physicochemical constraints and different cellular localizations. For example, "housekeeping" functions are under strong selection for optimizing their function in the normal habitat of the cell (usually making it fast and accurate under exponential growth conditions), whereas outer membrane proteins are often under selection for diversification (to explore a variety of substrates or to evade the immune system of hosts) (Finlay and Falkow 1997). Recently, it was found that genes presenting a large codon usage bias in both *E.*

*coli* and *Bacillus subtilis* code for a set of metabolically less costly amino acids (Akashi and Gojobori 2002). This was interpreted as advantageous in proteins translated at a high rate because it lowers the metabolic cost of translation. As a consequence, such proteins should be more evolutionary constrained and thus evolve more slowly.

In parallel to these studies, several conflicting reports have tried to take into account the role of essentiality in the rate of protein evolution. Early works proposed that proteins subject to the same type and level of functional constraints, but differing in terms of dispensability, should evolve at different rates because purifying selection would be more efficient on essential proteins (Wilson, Carlsson, and White 1977). However, a seminal analysis exploring a large set of mouse knock-out mutants showed that, when the tissue specificity of a gene is taken into account, there is no significant difference between the rates of evolution of essential and nonessential genes (Hurst and Smith 1999). Hirsh and Fraser (2001) further observed that essential genes do not evolve faster than nonessential genes in yeast. Quantification of the decrease in fitness associated with the inactivation of a gene is difficult because few systematic experimental data are available and because genes have different fitness effects in different environmental conditions. However, when dispensability was defined according to the loss of fitness associated with gene loss under exponential growth conditions, essentiality was reported to show a small but significant correlation with the rate of protein evolution (Hirsh and Fraser 2001).

Recently, two reports have further investigated the yeast data, strikingly reaching opposite conclusions. Pal, Papp, and Hurst (2003) analyzed protein substitution rates using three close relatives of *S. cerevisiae* and fitness data from whole-genome transcriptome data. The regression of dispensability on amino acid substitution rates, when controlled for expression level, was nonsignificant, suggesting that expression level is responsible for the small effect of dispensability on protein substitution rates. In contrast, Hirsh and Fraser (Hirsh 2003), using a different

Key words: protein evolution; substitution rates; eukaryotes; essentiality; expression levels; functional categories.

E-mail: erocha@abi.snv.jussieu.fr.

*Mol. Biol. Evol.* 21(1):108–116, 2004

DOI: 10.1093/molbev/msh004

*Molecular Biology and Evolution* vol. 21 no. 1

© Society for Molecular Biology and Evolution 2004; all rights reserved.

methodology to identify orthologs and a different set of transcriptome data, confirmed their previous conclusions, showing a significant correlation between dispensability and the nonsynonymous substitution rates, even when controlled for expression levels. In bacteria, the analysis of *E. coli* genes has shown that essential genes are more conserved than the bulk of the genes (Jordan et al. 2002). Further, when these authors inferred essentiality by homology in *Helicobacter pylori* and *Neisseria meningitidis*, the differences were also found to be significant, although smaller. Taken together, these observations suggest that essentiality is a determinant of amino acid substitution rates in bacterial proteins but not in higher eukaryotes. Yeast would be an intermediate case, where essentiality is not important but differential dispensability might be.

The differences between bacteria and yeast and the contradictory results obtained for the yeast data prompted us to revisit the determinants of amino acid substitution rates of bacterial proteins, using newly available data on essentiality and including all variables previously considered as separate factors. This was done in a multivariate analysis. The definition of the essential character of a gene is contingent to the set of experimental procedures aiming at such determination. Unfortunately, few genomes have been fully characterized in relation to essentiality: *S. cerevisiae* (Giaever et al. 2002), *B. subtilis* (Kobayashi et al. 2003), and *C. elegans* (Kamath et al. 2003). In the PEC database, information about *E. coli* gene essentiality has been extracted from the literature. This compilation is most useful, but, resting on evidence based on highly variable experimental set ups, it is not immune to the biases created by the lack of coordination of experiments developed with *E. coli*, especially because the large majority of these experiments were not specifically designed for uncovering lethal phenotypes. In both bacteria, the level of gene expression can be inferred from the bias resulting in optimization of codon usage for optimal growth under fast growth conditions (Ikemura 1981; Sharp and Li 1986; Andersson and Kurland 1990). Using the complete set of data available for gene inactivation in *B. subtilis* and the compilation available for *E. coli*, we use the complete sequences of related genomes to analyze the roles of the different factors on protein evolution, in particular with respect to essentiality and expressiveness.

## Materials and Methods

### Data

Data on the complete genomes of *B. subtilis* (GenBank accession number AL009126), *B. halodurans* (BA000004), *Oceanobacillus iheyensis* (BA000028), *E. coli* K12 (U00096), *Salmonella enterica* serovar Typhimurium (AE006468), and *Yersinia pestis* (AL590842) were taken from GenBank genomes (<ftp://ftp.ncbi.nih.gov/genomes/Bacteria/>). Data on the essential genes of *B. subtilis* were taken from the literature (Kobayashi et al. 2003). They include a list of 277 lethal genes in the genome (~4,100 genes). This data set corresponds to the coordinated analysis of extensive gene knock-out experiments in *B. subtilis*. Data on the essential genes of *E. coli*

were taken from the PEC database (<http://www.shigen.nig.ac.jp/ecoli/pec/>), which included information on 203 essential genes out of the approximately 4,300 genes of the *E. coli* K12 genome.

### Functional Classification

The information on the functional classification of genes in *B. subtilis* and *E. coli* K12 was taken from the corresponding genome sequencing papers (Blattner et al. 1997; Kunst et al. 1997). Proteins were classed into “envelope,” “metabolism,” “information,” and “others & UFO (unknown function ORF).” In *B. subtilis*, the class “envelope” included every category under “cell envelope and cellular processes,” excluding “sporulation,” “germination,” and “competence.” The class “metabolism” included everything under the category “intermediary metabolism.” The class “information” included everything under the category “information pathways.” The remaining elements and the unknown function CDSs (UFO) were classified as “others & UFO.” We made small changes to the categories of the *E. coli* sequencing paper to render them as comparable with the ones of *B. subtilis* as possible. Thus, “metabolism” and “energy metabolism” were merged; “cell structure,” “membrane proteins,” and “transport proteins” were merged into “envelope”; and the “information” class was built from merging of “replication,” “transcription,” and “translation.” The remaining elements and the UFO were classified as “others & UFO.”

### Phylogenetic Analysis

It is known from the literature than among the three enterobacteria, *E. coli* and *Salmonella* are monophyletic relative to *Yersinia* (Neidhardt et al. 1996). Because we failed to find in the literature the phylogenetic relation between the two *Bacillus* sp. and *O. iheyensis*, we derived it using the translation elongation factor EFTu as a phylogenetic marker. We used *S. aureus* as an outgroup to be able to root the tree and precisely determine which couple is monophyletic relative to the third genome. The EFTu protein sequences were aligned and then back-translated into DNA to build a matrix of distances using maximum likelihood under the HKY85 model (Hasegawa, Kishino, and Yano 1985) in Tree-Puzzle (Schmidt et al. 2002), using the Gamma correction. A tree was then built using the distance matrix and the BIONJ program (Gascuel 1997). The robustness of branches was tested by 1,000 bootstrap experiments, using SEQBOOT and CONSENSE, from the PHYLIP package (Felsenstein 1993). This analysis indicated with large confidence (999/1,000 bootstraps) that the two *Bacillus* sp. are monophyletic relative to *O. iheyensis*. The two-way comparison of similarity between all orthologs of *B. subtilis* with the two other genomes confirmed the topology of the tree, because orthologs in *B. subtilis* and *B. halodurans* show systematically higher similarity between themselves than with *O. iheyensis* (paired *t*-test,  $P < 0.001$ ).

## Analysis of Orthology

Orthologs were identified as reciprocal best hits (Tatusov and Koonin 1997) (using a global alignment where the gaps on the edges of the largest sequence are ignored) with at least 50% similarity in amino acid sequence and less than 20% difference in protein length. We identified the orthologs between every pair of each of the two triplets (i.e., between each pair of the three *Bacillus/Oceanobacillus* genomes and each pair of the *E. coli/Salmonella/Yersinia* genomes). Then, in each triplet, we rejected all orthologs that were not simultaneously present in the three genomes or that gave different correspondences in different comparisons (e.g., the ortholog resulting from the comparison between *B. subtilis* with *B. halodurans* and *B. subtilis* with *O. iheyensis* was not the same as the one obtained by the comparison of *B. halodurans* with *O. iheyensis*).

Because we know the phylogenetic tree describing the evolutionary history of these bacteria, we made a further analysis to remove potential false orthologs that could arise from horizontal transfer or differential gene deletion. Consider the three genomes A, B, and C, where A and B are monophyletic (i.e., *B. subtilis* and *B. halodurans* in one group and *E. coli* and *Salmonella* in the other group). Consider  $\alpha$ ,  $\beta$ , and  $\gamma$  a triplet of orthologs of these genomes. One would expect the similarity between orthologs in the three genomes ( $S_{\alpha,\beta}$ ,  $S_{\alpha,\gamma}$ ,  $S_{\beta,\gamma}$ ) to obey the relationship  $S_{\alpha,\beta} > S_{\alpha,\gamma} \sim S_{\beta,\gamma}$ , and we use this information to further filter our data. We allow for a small interval of tolerance (5%) in the first inequality, and thus we eliminate all triplets where

$$\frac{S_{\beta,\gamma} - S_{\alpha,\beta}}{S_{\beta,\gamma} + S_{\alpha,\beta}} > 0.05.$$

## Substitution Rates

The rates of synonymous ( $d_S$  and  $K_S$ ) and non-synonymous ( $d_N$  and  $K_A$ ) substitutions were computed following Yang's definition (Yang and Nielsen 2000) ( $d_N$  and  $d_S$ ) using PAML and following Li's definitions (Li 1993) ( $K_A$  and  $K_S$ ) using Jadis (Gonçalves et al. 1999). All results presented in the article refer to  $d_N$  and  $d_S$ . The values of  $K_A$  and  $K_S$  were only used for verification.

## CAI Calculations

Codon adaptation index (CAI) values were computed using the EMBOSS package (<http://www.uk.embnet.org/Software/EMBOSS>). The reference values of codon usage in highly expressed genes were computed using the ribosomal proteins as markers of the translation machinery that is the largest processing machinery under exponential growth conditions. When possible, we preferred to take into consideration the quantitative measure of CAI, but when a qualitative variable seemed useful, we classed genes as highly expressed and non-highly expressed. A gene was regarded as potentially highly expressed if its CAI was among the 20% highest values of the genome. Variations around the value of 20% (from 10% to 30%) did not significantly alter the results (data not shown).

## Metabolic Cost of Amino Acids

The metabolic cost of each amino acid was computed using publicly available data (Akashi and Gojobori 2002). This cost takes into account the metabolic pathways leading to amino acid biosynthesis both in *E. coli* and *B. subtilis*. Energetic costs are converted to a single currency of  $\sim P$ , based on a proportion of two  $\sim P$ s for one H (in NADH, NADHP, and FADH<sub>2</sub>). Each protein was then associated with an average amino acid cost per residue.

Given the sequence similarities between the genomes of the two groups, we have computed  $d_N$  and  $K_A$  values between *B. subtilis* and *B. halodurans* (henceforth named the *B. subtilis* group) and  $d_N$ ,  $K_A$ , and  $d_S$  and  $K_S$  values between *E. coli* K12 and *S. enterica* serovar Typhimurium (henceforth named the *E. coli* group). The orthologs of the first group were classed into functional categories, CAI, and essentiality according to the information available for *B. subtilis*. The orthologs of the second group were classed according to the information available for *E. coli*. Not all *E. coli* genes are yet classed as essential or nonessential. To simplify the analysis, we removed all these genes. Thus, the *B. subtilis* data set includes 1,258 orthologs, and the *E. coli* data set includes 1,364 orthologs.

## Results

### Correlation Between Variables

We started our work by analyzing the correlation of nonsynonymous substitutions rate with the other variables. For simplicity, and because both methods for the determination of synonymous and nonsynonymous substitution rates give qualitatively similar results, only the analyses with  $d_N$  and  $d_S$  will be shown (see Supplementary Material online for the analyses with  $K_A$  and  $K_S$ ). Nonsynonymous substitutions are not independent of any of the other variables (figs. 1 and 2). In the *Bacillus* group, essential genes show a median  $d_N$  of 0.18 against a median of 0.30 for the remaining genes ( $P < 0.001$ , Wilcoxon test). Among the four functional categories, differences were neither significant between "metabolism" and "information" nor between "envelope" and "others & UFO" ( $P < 0.01$ , Tukey-Kramer HSD tests). On the other hand, these two superclasses were found to be significantly different ( $P < 0.01$ , Wilcoxon test). CAI is negatively correlated with  $d_N$  (Spearman's  $\rho = -0.46$ ,  $P < 0.001$ ). The division of genes in highly expressed genes and non-highly expressed genes confirmed this analysis showing a median  $d_N$  of 0.14 for the former and of 0.31 for the latter ( $P < 0.001$ , Wilcoxon test). The metabolic cost of amino acids is positively correlated with  $d_N$  (Spearman's  $\rho = 0.23$ ,  $P < 0.001$  [fig. 2]). We found qualitatively similar results for the correlation of the different variables with  $d_N$  in *E. coli*. The major quantitative difference concerned the variable amino acid metabolic cost, which is less important than in *B. subtilis* (Spearman's  $\rho = 0.061$ ,  $P < 0.01$ ).

All variables are thus significantly correlated with  $d_N$  both in *B. subtilis* and in *E. coli*. In fact, they are also correlated among themselves. In *B. subtilis*, the distribution of the functional categories is quite different between

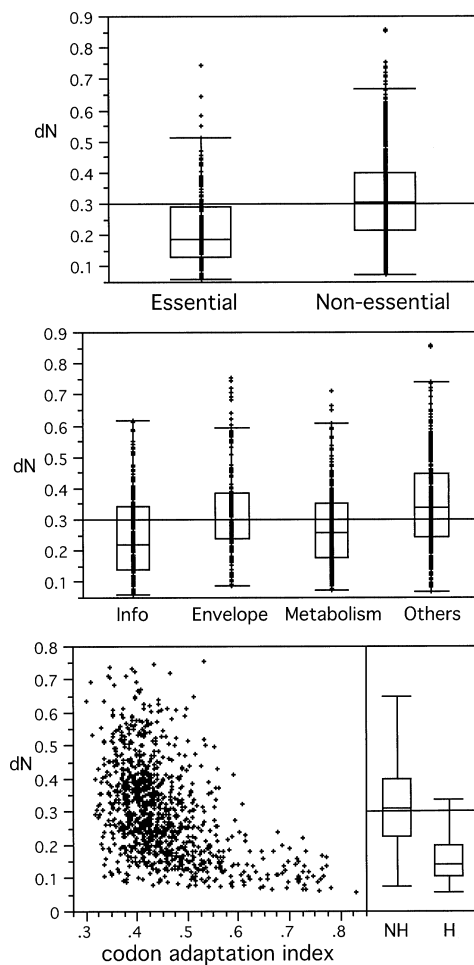


FIG. 1.—Distribution of the rate of nonsynonymous substitution ( $d_N$ ) computed between the genes of *B. subtilis* and *B. halodurans* in terms of the different variables: essentiality, functional category, and expression (using CAI). H indicates highly expressed genes; NH indicates non-highly expressed genes. The edges of the boxes indicate the upper and lower quartiles. The line at the center of the box indicates the median, and the edges of the upper/lower whiskers represent the limits of 1.5 times the upper/lower interquartile ranges. The horizontal lines indicate the grand mean.

essential and nonessential genes, with a significant overrepresentation of “information” genes among essential and highly expressed genes (fig. 3A). This is because of the weight of the proteins involved in translation, which are essential and highly expressed for the most part. On the other hand, the category “others & UFO” is significantly underrepresented in the sets of highly expressed and essential genes. Similar results were obtained when analyzing the *E. coli* data (fig. 3B). In this case, many of the genes, especially in the class “others & UFO,” have not yet been tested for an essential phenotype. The metabolic cost of amino acids is known from the literature to be correlated with codon usage (Akashi and Gojobori 2002). Because all variables are correlated with each other and with  $d_N$ , conclusions drawn on the analysis of each separate variable may include effects that are caused by the other variables. We used multivariate regressions to disentangle these effects (Draper and Smith 1998).

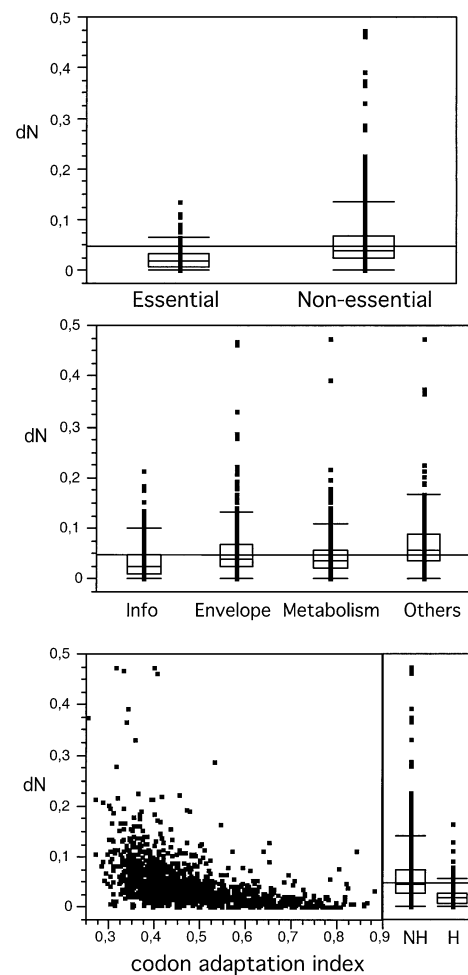


FIG. 2.—Distribution of the rate of nonsynonymous substitution ( $d_N$ ) computed between the genes of *E. coli* and *S. enterica* in terms of the variables essentiality, functional category, and expression (using CAI). H indicates highly expressed genes; NH indicates non-highly expressed genes. The edges of the boxes indicate the upper and lower quartiles. The line at the center of the box indicates the median, and the edges of the upper/lower whiskers represent the limits of 1.5 times the upper/lower interquartile ranges. The horizontal lines indicate the grand mean.

#### Multiple Regressions on the *Bacillus* Data

Given the distribution of  $d_N$  in function of the CAI values (fig. 1), we transformed the data to apply the standard procedures of linear multiple regression. The Box-Cox method indicated that the data should be transformed with a logarithm (Draper and Smith 1998), and all regressions were done using a logarithmic transformation on  $d_N$ . The basic statistical fundamentals of ANOVA and regression techniques are the same. In general, ANOVA is used to analyze the role of nominal/ordinal variables and regression is used to analyze continuous variables. Here, we use multiple regressions that include both continuous and discrete variables, allowing the mixing of both types of variables (Sokal 1981; Zar 1996). In this case, discrete variables are coded under the form of “dummy variables.” The exact value of the coding ( $\pm 1$  in this work) is important for the

interpretation of the regression equations but not for the analysis of the relative importance of the variables in the regression.

Then we made forward and backward stepwise regressions to inspect whether all variables provide significant information. Both methods led to similar results, and thus only the results of the forward stepwise regression method will be shown here. In a forward stepwise regression, one begins with the smallest possible regression model, the one including the variable showing the highest  $R^2$ . The model is then built up by successively adding the most significant variables (Draper and Smith 1998). This analysis confirmed that functional categories are better analyzed if pooled together into two groups, one with “information” and “metabolism” and the other with “envelope” and “others & UFO” (data not shown). The other combinations of categories provided no significant contributions to the regression (even at  $P < 0.2$ ). As a consequence, functional categories were pooled into those two superclasses. We tested whether interaction terms might be significant. In all cases, these analyses indicated that such terms provided no significant contribution to the regression fit (data not shown). At this stage we could define the model to be tested. It has the form:

$$\log(dN) = \beta_0 + \beta_{CAI}X_{CAI} + \beta_{cost} \log(X_{cost}) + \beta_{cat}X_{cat} + \beta_{ess}X_{ess},$$

where essentiality ( $X_{ess}$ ) and functional category ( $X_{cat}$ ) are dichotomous nominal variables (taking values +1/-1) and CAI ( $X_{CAI}$ ), and metabolic cost ( $X_{cost}$ ) are continuous variables. The first element integrated in the equation by forward stepwise regression is the CAI (table 1). The inclusion of this variable leads to a coefficient of determination that corresponds to 91% of the global  $R^2$  (which is 0.322,  $P < 0.001$ ). The second variable entering in the regression is the functional category (+8% of the  $R^2$ ) and then the essential character of the gene (+1% of the  $R^2$ ). The metabolic cost, which has a very minor role in its regression with  $d_N$  ( $R^2 = 0.048$ ), does not contribute significantly to the multiple stepwise regression.

This analysis suggests that CAI is the most relevant determinant of the rate of protein evolution. However, because the variables are correlated, the inclusion of CAI also leads to the inclusion of part of the information present in the other variables, via the correlation of these variables with CAI. To test if this could affect the previous observation, we performed separate linear regressions of  $d_N$  on each of the variables. They confirm our analysis (table 1). Regression of  $\log d_N$  on CAI is highly significant ( $R^2 = 0.289$ ,  $P < 0.0001$ ), whereas regression on the functional categories ( $R^2 = 0.084$ ,  $P < 0.0001$ ) and regression on essentiality ( $R^2 = 0.097$ ,  $P = 0.04$ ) are much less important. We have also made an additional analysis where we code CAI as a discrete variable (see *Materials and Methods*). All three most-significant variables are discrete in this case. Even though this leads to loss of information in the variable measuring codon usage bias, CAI remains responsible for 74% of the total  $R^2$  (see Supplementary Material online).

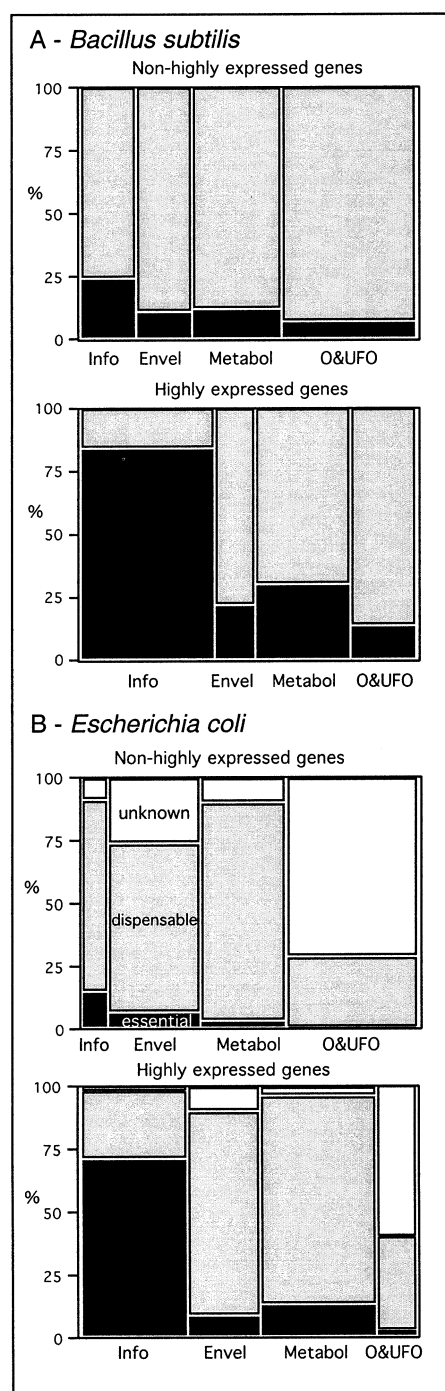


FIG. 3.—Distribution of essential (black areas), nonessential (gray areas), and noncharacterized genes (white areas) according to expression levels and functional categories in *B. subtilis* (A) and *E. coli* (B). The surface of each class is proportional to its relative frequency in the data.

#### Multiple Regression of the *E. coli* Data

As for the *Bacillus* group, the preliminary analysis of the *E. coli* data suggested a log transformation of  $d_N$ . We also merged together the four functional classes into only two categories, one related with “information” and

**Table 1**  
**Results of the Regression Analyses on the 1,258 Genes of *B. subtilis* and the 1,364 Genes of *E. coli*, Using Log  $d_N$  and Log  $d_S$**

Parameters	Full Multiple Regression <sup>a</sup>		Stepwise Regression <sup>b</sup>		
	Estimate	<i>P</i>	Insertion Order <sup>c</sup>	$R^2$ (%) <sup>d</sup>	$R^{2*e}$
<i>B. subtilis</i> ( $R^2 = 0.322$ , $P < 0.0001$ )					
Intercept	$-0.526 \pm 0.375$	—	—	—	—
CAI	$-1.296 \pm 0.083$	<0.0001	1	91	0.289
Category	$-0.034 \pm 0.006$	<0.0001	2	99	0.084
Essentiality	$-0.019 \pm 0.009$	0.038	3	100	0.097
Log cost	$0.375 \pm 0.266$	0.159	NS	—	0.048
<i>E. coli</i> ( $R^2 = 0.307$ , $P < 0.0001$ )					
Intercept	$0.704 \pm 0.508$	—	—	—	—
CAI	$-1.732 \pm 0.088$	<0.0001	1	94	0.287
Category	$-0.022 \pm 0.009$	0.016	4	99	0.033
Essentiality	$-0.066 \pm 0.013$	<0.0001	2	98	0.080
Log cost	$-1.016 \pm 0.365$	0.005	3	100	0.002
<i>E. coli</i> , including $d_S$ ( $R^2 = 0.401$ , $P < 0.0001$ )					
Intercept	$0.024 \pm 0.475$	—	—	—	—
Log $d_S$	$0.481 \pm 0.033$	<0.0001	1	85	0.343
CAI	$-1.002 \pm 0.096$	<0.0001	2	99	0.287
Category	$-0.016 \pm 0.008$	0.054	3	99.5	0.033
Essentiality	$-0.017 \pm 0.013$	0.191	NS	—	0.080
Log cost	$-0.752 \pm 0.340$	0.027	4	100	0.002

NOTE.—Log  $d_N$  and Log  $d_S$  as defined by Yang and Nielsen (2000). NS indicates not significant.

<sup>a</sup> Estimates of the parameters of the multiple regressions, their standard deviations, and their significance.

<sup>b</sup> Results concerning the forward stepwise regressions.

<sup>c</sup> Order of introduction of variables in the stepwise regression (corresponds to their relative importance in the global fit).

<sup>d</sup> Percentage of the global  $R^2$  obtained when the variable is included in the stepwise regression (in the order given by the previous column). Using the *B. subtilis* data as an example, first CAI is included (insertion order = 1), leading to a regression with an  $R^2$  corresponding to 91% of the total  $R^2$ . Then the variable categories are included (insertion order = 2), and  $R^2$  increases to 99% of the total (thus the step adds 8% of the  $R^2$ ). The inclusion of essentiality (insertion order = 3) adds the remaining 1% to the total  $R^2$ .

<sup>e</sup> Coefficient of determination obtained in the simple linear regressions of  $d_N$  with each variable (e.g., 0.289 for the simple regression of  $d_N$  on CAI for the *B. subtilis* data).

“metabolism” and the other including the remaining genes. The simple linear regressions, the stepwise regression, and the full multiple regression all revealed that *E. coli* shares the major characteristics of *Bacillus*. In the stepwise regression, the introduction of CAI leads to 94% of the total  $R^2$  (which is 0.307, similar to *Bacillus*). The major difference between this analysis and the one of *B. subtilis* concerns the smaller role of the variable functional category, which only ranks fourth ( $R^2$  of 0.033 versus 0.084 in the *Bacillus*). This may be a consequence of two factors. First, the classification schemes used in *E. coli* are, despite our best efforts, slightly different from the ones of *B. subtilis*. Second, the genes that are not classified regarding their essentiality are unevenly distributed among functional classes (fig. 3).

The simple regressions, to test whether the relatively larger importance of CAI indicated by the stepwise regression is not an effect of the correlation between the variables, confirm the important correlation of CAI with the nonsynonymous substitution rate. The regression of CAI shows an  $R^2$  of 0.287, whereas the regressions of the other variables never exceed an  $R^2$  of 0.08 (table 1). The metabolic cost of amino acids is the third variable to enter the stepwise regression, for which it contributes very little (1%). The simple regression of this variable with  $d_N$  shows an  $R^2$  of 0.002, indicating that the metabolic cost of amino acids explains at most 2% of the total variance. Finally, the

regression of the full model shows the same trends of the one of *Bacillus* (table 1).

#### Including $d_S$ in the Analysis of the *E. coli* Group

Although the comparison of the *E. coli* group is hampered by the lack of an exhaustive study of essential phenotypes, the divergence of these genomes allows for the analysis of the rates of synonymous substitution ( $d_S$ ) between *E. coli* and *S. enterica* (Sharp 1991; Berg 1999). Thus, our first concern was to quantify the correlations between CAI,  $d_N$ ,  $d_S$  (both properly transformed), and the metabolic cost of amino acids. CAI shows similar values of pairwise correlation with log  $d_N$  and log  $d_S$  (table 2). This suggests that the correlation between  $d_N$  and CAI is not simply the result of the correlation of both variables with  $d_S$ . We thus computed the partial correlations between the variables; that is, the correlation between each pair of variables while holding constant the value of the other variables (Zar 1996). All partial correlations between CAI, log  $d_N$ , and log  $d_S$  are statistically significant, confirming that  $d_N$  correlates with CAI independently of  $d_S$  (table 2). The metabolic cost of amino acids correlates poorly with both  $d_S$  and  $d_N$  and negatively with CAI.

We then proceeded to test if the introduction of  $d_S$  in the linear model could change significantly the relative

importance of CAI, essentiality, and functional category. Under these circumstances, the model becomes:

$$\log(dN) = \beta_0 + \beta_{CAI}X_{CAI} + \beta_{cost} \log(X_{cost}) + \beta_{cat}X_{cat} + \beta_{ess}X_{ess} + \beta_{dS} \log(X_{dS}).$$

The regressions show that the introduction of  $d_S$  renders essentiality uninformative, whereas CAI still contributes significantly to the model (table 1). Thus, although there is a significant correlation between  $d_N$  and  $d_S$ , the introduction of  $d_S$  in the model confirms the main conclusion of our work: that essentiality, functional categories, and metabolic cost are, at most, minor determinants of the rate of protein evolution. Also, it shows that CAI and  $d_N$  correlate in an important way and independently of  $d_S$ .

## Discussion

### The Minor Roles of Functional Category and Metabolic Cost

The functional category of proteins is among the variables that we found to play a minor (if any) role in the rate of protein evolution. Metabolic and information proteins evolve at similar rates but slower than the other proteins. These functional categories are very broad and the variance in each of them does not exclude that particular categories may be more or less conserved. The analysis of genomes closely related to *E. coli* and *B. subtilis* allowed us to concentrate on genes most probably shared since speciation, eliminating interference from horizontal transfer. As such, it is interesting to analyze our data in the light of the “complexity” hypothesis (Jain, Rivera, and Lake 1999). This hypothesis states that macroevolutionary trends of information proteins to evolve slower than metabolic proteins are caused by lack of horizontal transfer among the former. Indeed, when we eliminated horizontal transfer, we did not find this class to evolve significantly slower than the one of metabolic proteins.

Genes with high codon usage biases tend to use metabolically less expensive amino acids (Akashi and Gojobori 2002). The simple regressions of this cost on  $d_N$  only explain 4.8% (*B. subtilis*) and 0.2% (*E. coli*) of the total variance. Such contribution becomes insignificant when the other variables are included. Therefore, the putative selection pressure for metabolic efficiency does not change the rate of nonsynonymous substitutions. This is also true for the rate of synonymous substitutions in *E. coli* because the partial correlation between the two variables is very low ( $<0.02$ ). This is unexpected, because a selective pressure acting on the usage of less expensive amino acids should lead to a smaller probability of fixation of synonymous and nonsynonymous substitutions at these sites. Further work will be necessary to establish whether such selection pressure exists.

### Expression Levels and the Rate of Protein Evolution

CAI is a good measure for the level of gene expression under exponential growth in fast-growing organisms, as evidenced by the correlation of the

**Table 2**  
Pearson's Correlations and Partial Correlations Between CAI, Log  $d_S$ , Log  $d_N$ , and the Metabolic Cost of Proteins per Amino Acid for the *E. coli* Data

	Log $d_N$	Log $d_S$	CAI	Log Cost
Log $d_N$	—	0.586	−0.536	0.047
Log $d_S$	0.389	—	−0.599	0.083
CAI	−0.291	−0.412	—	−0.196
Log cost	−0.058	−0.018	−0.192	—

NOTE.—Pearson's correlations: upper diagonal. Partial correlations: lower diagonal.

frequency of optimal codons with the concentration of the cognate tRNA (Dong, Nilsson, and Kurland 1996) and of the values of CAI with the corresponding mRNA (Coghlan and Wolfe 2000) and protein (Futcher et al. 1999) concentrations. In fact, codon usage bias relates intimately with high expression levels because genes are under selection pressure for the use, for a given amino acid, of the codon(s) corresponding to the most abundant tRNAs (Ikemura 1981). As a result, higher codon usage bias resulting from the selection of optimal codons leads to a higher number of deleterious synonymous substitutions and thus to lower synonymous substitution rates (Sharp 1991). Translation accuracy may also play a role in the establishment of codon usage biases (Akashi 1994). Under the accuracy hypothesis, codon usage also reflects selection for the use of codons that favor lower levels of mistranslation. Proteins for which a larger fraction of substitutions lead to a significant decrease of function (thus lower nonsynonymous substitution rates) would also strongly select “accurate” codons (thus lower synonymous substitution rates and higher CAI). This could justify the correlation found between the rates of synonymous and nonsynonymous substitutions (Li, Wu, and Luo 1985; Mouchiroud, Gautier, and Bernardi 1995).

Lobry and Gautier (1994) suggested a different cause for the correlation of nonsynonymous substitution rates with CAI. They found that highly expressed genes have an amino acid composition that matches the most abundant tRNAs. Under these circumstances, they suggest that highly expressed genes reduce the diversity of amino acid choices to increase translation efficiency. The major difference between these hypotheses relies on the sense of the relationship between the rate of protein evolution and codon usage bias. The accuracy hypothesis indicates that pressure for protein sequence conservation leads to codon usage bias, whereas the latter hypothesis suggests the inverse.

Double mutations, because they typically involve a synonymous and a nonsynonymous substitution (Averof et al. 2000), could lead to a correlation between  $d_N$ ,  $d_S$ , and CAI. However, several recent works indicate a negligible role for double mutations in genomes (Smith and Hurst 1999; Smith, Webster, and Ellegren 2003). Further, the removal of double substitutions does not eliminate the correlation between synonymous and nonsynonymous substitution rates (Mouchiroud, Gautier, and Bernardi 1995) nor between nonsynonymous substitution rates and the expression level (Pal, Papp, and Hurst 2001). Finally,

our results indicate that CAI correlates with nonsynonymous substitution rates almost as strongly as with synonymous substitution rates. This suggests a direct link between expression level and the rate of protein evolution.

One could also suppose that protein expression levels relate directly to the rate of protein evolution because substitutions will be more deleterious in proteins that have a larger impact on fitness and that such an impact is likely to correlate to the proteins expression level. As an example, let us consider two proteins with metabolic nonessential functions and very different expression levels, for which one deleterious mutation renders their biochemical function 5% less efficient. The impact of this efficiency loss on the cell's fitness will be the product of the loss of their biochemical efficiency by the relative weight of the corresponding reactions in the cell metabolism (Hartl, Dykhuizen, and Dean 1985). Under these circumstances, one would expect the same relative efficiency loss to have a larger impact on the cell's metabolism, and thus on the cell fitness, for highly expressed genes. Naturally, many other factors interfere with the rate of protein evolution. Some functions have an importance to the cell's fitness that does not correspond to the expression levels of the corresponding genes. DNA replication, for example, is likely to be under important purifying selection despite the typically low expression levels of the DNA polymerase. Also, expression levels depend on physiological conditions, and some proteins are highly expressed under some conditions and not expressed at all under other conditions. This problem concerns all the hypotheses so far envisaged.

### The Role of Essentiality

It has been suggested that essentiality explains a significant part of the overall variance of nonsynonymous substitution rates in bacteria (Jordan et al. 2002). However, this analysis did not control for the effects of expression levels. When this effect is taken into account, essentiality explains very little of the remaining variance. When the regression of the rate of nonsynonymous substitutions in function of essentiality is controlled for expression levels, we find very low  $R^2$  (0.007 for *B. subtilis* and 0.012 for *E. coli*). Thus, under these circumstances, essentiality explains approximately 1% of the variance in both bacteria. This strongly suggests that essentiality and nonsynonymous substitution rates are related fundamentally via the correlation of both variables with CAI, and that the lack of role of essentiality in the rate of protein evolution in bacteria is the same as in eukaryotes. We have pointed out above the contradiction between the different analyses of the yeast data (Hirsh 2003; Pal, Papp, and Hurst 2003). Hirsh and Fraser's argument is based on the standard population genetics reasoning that the probability of fixation of a deleterious substitution is expected to be high only for proteins whose effect in fitness is small (Hirsh and Fraser 2001). According to this view, one could expect essential proteins to contain a larger number of sites under strong selection, which would result in lower nonsynonymous substitution rates. Our observations indicate otherwise. The observa-

tion that essentiality, after controlling for expression, explains 1% of the overall variance is very similar to the one found in yeast by Pal, Papp, and Hurst (2003), using dispensability, and quite smaller than the one found by Hirsh and Fraser (Hirsh 2003). Unfortunately, one cannot precisely compare our results with the yeast data, because systematic dispensability data is currently unavailable in bacteria. Yet, the coincidence between our results and the ones of Pal, Papp, and Hurst (2003) suggests that the role of essentiality is minor both in bacteria and in eukaryotes.

### Supplementary Material

Supplementary Material is available on the journal's Web site and at <http://www.abi.snv.jussieu.fr/~erocha/protevol/>.

### Acknowledgments

We are grateful to Isabelle Gonçalves and Carmen Bessa-Gomes for comments on the manuscript. In addition to contribution of the scientists making the Stanislas Noria network on conceptual biology, this work is based on the core experimental work performed by the consortium of *Bacillus subtilis* scientists who, both in Europe and in Japan, deciphered the genome sequence and disrupted and studied all the genes of the genome. We wish here to express our thanks to all these scientists and, in particular, to those who made the enterprise happen, S. D. Ehrlich, F. Kunst, N. Ogasawara, and H. Yoshikawa.

### Literature Cited

- Akashi, H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136**:927–935.
- Akashi, H., and T. Gojobori. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc. Natl. Acad. Sci. USA* **99**: 3695–3670.
- Andersson, S. G. E., and C. G. Kurland. 1990. Codon preferences in free-living microorganisms. *Microbiol. Rev.* **54**:198–210.
- Averof, M., A. Rokas, K. H. Wolfe, and P. M. Sharp. 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* **287**:1283–1286.
- Berg, O. G. 1999. Synonymous nucleotide divergence and saturation: effects of site-specific variations in codon bias and mutation rates. *J. Mol. Evol.* **48**:398–407.
- Blattner, F. R., G. P. III, C. A. Bloch et al. (17 co-authors). 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**:1453–1461.
- Coghlan, A., and K. H. Wolfe. 2000. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast* **16**:1131–1145.
- Dickerson, R. E. 1971. The structure of cytochrome c and the rates of molecular evolution. *J. Mol. Evol.* **1**:26–45.
- Dong, H., L. Nilsson, and C. G. Kurland. 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J. Mol. Biol.* **260**:649–663.
- Draper, N. R., and H. Smith. 1998. Applied regression analysis. John Wiley & Sons, New York.
- Felsenstein, J. 1993. PHYLIP (phylogeny inference package). Version 3.6a. Distributed by the author, Department of Genetics, University of Washington, Seattle.

- Finlay, B. B., and S. Falkow. 1997. Common themes in microbial pathogenicity revisited. *Microbiol. Mol. Biol. Rev.* **61**:136–169.
- Futcher, B., G. I. Latter, P. Monardo, C. S. McLaughlin, and J. I. Garrels. 1999. A sampling of the yeast proteome. *Mol. Cell Biol.* **19**:7357–7368.
- Gascuel, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**:685–695.
- Giaever, G., A. M. Chu, C. Connelly et al. (73 co-authors). 2002. Functional profiling of the *Saccharomyces cerevisiae* genomes. *Nature* **418**:387–391.
- Gonçalves, I., M. Robinson, G. Perriere, and D. Mouchiroud. 1999. JaDis: computing distances between nucleic acid sequences. *Bioinformatics* **15**:424–425.
- Hartl, D. L., D. E. Dykhuizen, and A. M. Dean. 1985. Limits of adaptation: the evolution of selective neutrality. *Genetics* **111**:655–674.
- Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**:160–174.
- Hirsh, A. E. 2003. Rate of evolution and gene dispensability—reply. *Nature* **421**:497–498.
- Hirsh, A. E., and H. B. Fraser. 2001. Protein dispensability and rate of evolution. *Nature* **411**:1046–1049.
- Hurst, L. D., and N. G. Smith. 1999. Do essential genes evolve slowly? *Curr. Biol.* **9**:747–750.
- Ikemura, T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* **146**:1–21.
- Jain, R., M. C. Rivera, and J. A. Lake. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. USA* **96**:3801–3806.
- Jordan, I. K., I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* **12**:962–968.
- Kamath, R. S., A. G. Fraser, Y. Dong et al. (13 co-authors). 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**:231–237.
- Kobayashi, K., S. D. Ehrlich, A. Albertini et al. (99 co-authors). 2003. Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. USA* **100**:4678–4683.
- Kunst, F., N. Ogasawara, I. Moszer et al. (151 co-authors). 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**:249–256.
- Li, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36**:96–69.
- Li, W.-H., C.-I. Wu, and C. C. Luo. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide codon changes. *Mol. Biol. Evol.* **2**:150–174.
- Lobry, J. R., and C. Gautier. 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Res.* **22**:3174–3180.
- Mouchiroud, D., C. Gautier, and G. Bernardi. 1995. Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of non-synonymous substitutions. *J. Mol. Evol.* **40**:107–113.
- Nei, M. 2000. *Molecular phylogenetics and evolution*. Sinauer Press, Sunderland, Mass.
- Neidhardt, F., R. Curtiss, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger. 1996. *Escherichia coli* and *Salmonella*: cellular and molecular biology. ASM Press, Washington, DC.
- Pal, C., B. Papp, and L. D. Hurst. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* **158**:927–931.
- . 2003. Rate of evolution and gene dispensability. *Nature* **421**:496–497.
- Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**:502–504.
- Sharp, P. M. 1991. Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position and concerted evolution. *J. Mol. Evol.* **33**:23–33.
- Sharp, P. M., and W.-H. Li. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **24**:28–38.
- Smith, N. G., and L. D. Hurst. 1999. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* **153**:1395–1402.
- Smith, N. G., M. T. Webster, and H. Ellegren. 2003. A low rate of synonymous double-nucleotide mutations in Primates. *Mol. Biol. Evol.* **20**:47–53.
- Sokal, R. R. 1981. *Biometry*. W. H. Freeman, New York.
- Tatusov, R. L., and E. V. Koonin. 1997. A genomic perspective of protein families. *Science* **278**:631–637.
- Wilson, A. C., S. S. Carlsson, and T. J. White. 1977. Biochemical evolution. *Annu. Rev. Biochem.* **46**:573–639.
- Yang, Z., and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**:32–43.
- Zar, J. H. 1996. *Biostatistical analysis*. Prentice Hall, New Jersey.

Geoffrey McFadden, Associate Editor

Accepted August 15, 2003