Combining Models of Protein Translation and Population Genetics to Predict Protein Production Rates from Codon Usage Patterns

Michael A. Gilchrist

Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville

Genes are often biased in their codon usage. The degree of bias displayed often changes with expression level and intragenic position. Numerous indices, such as the codon adaptation index, have been developed to measure this bias. Although the expression level of a gene and index values are correlated, the heuristic nature of these metrics limits their ability to explain this relationship. As an alternative approach, this study integrates mechanistic models of cellular and population processes in a nested manner to develop a stochastic evolutionary model of a protein's production rate (SEMPPR). SEMPPR assumes that the evolution of codon bias is driven by selection to reduce the cost of nonsense errors and that this selection is counteracted by mutation and drift. Through the application of Bayes' theorem, SEMPPR generates a posterior probability distribution for the protein production rate of a given gene. Conceptually, SEMPPR's predictions are based on the degree of adaptation to reduce the cost of nonsense errors observed in the codon usage pattern of the gene. As an illustration, SEMPPR was parameterized using the *Saccharomyces cerevisiae* genome and its predictions tested using available empirical data. The results indicate that SEMPPR's predictions are as reliable index based ones. In addition, SEMPPR's output is more easily interpreted and its predictions could be improved through refinements of the models upon which it is built.

Introduction

Codon usage bias is the selective (i.e., nonuniform) usage of particular synonymous codons within a genetic sequence (Ikemura 1981; Bennetzen and Hall 1982; Sharp and Li 1987). The nature and strength of codon bias varies between species as well as within a genome and across tissues. Correlating protein or mRNA abundance to codon bias has been the subject of numerous studies (Gouy and Gautier 1982; Ikemura 1985; Gygi et al. 1999; Greenbaum et al. 2003). Such approaches take a descriptive approach in linking codon usage and abundance. In contrast, this study demonstrates how codon usage bias and the average protein production rate of a gene can be linked mechanistically. This is done through the development of a stochastic evolutionary model of a protein's production rate (SEMPPR). SEMPPR is built by integrating nested models of intracellular and population-level processes together and applying Bayes' theorem.

Although it is well accepted that codon bias is positively correlated with gene expression level, the importance of different possible sources of selection is still unclear (Bulmer 1991; Kurland 1992; Akashi 1994, 1995; Eyre-Walker 1996; Akashi and Eyre-Walker 1998; Comeron and Kreitman 2002; Elf et al. 2003; Kliman et al. 2003; Qin et al. 2004; Gilchrist and Wagner 2006; Plotkin et al. 2006). One hypothesis is that codon bias results from selection to minimize the probability and associated costs of nonsense errors (i.e., the premature termination of protein translation).

A major cost of a nonsense error is the amount of energy invested into assembling the incomplete polypeptide (Bulmer 1991; Kurland 1992; Eyre-Walker 1996). Because this cost is related to the length of the nascent peptide when the error occurs, selection on codon usage against nonsense errors should increase with codon position along a sequence (Bulmer 1991; Kurland 1992; Qin et al. 2004; Gilchrist and Wagner 2006). This, in turn, leads to the prediction of increasing codon bias with codon position. The prediction of increasing bias with position is unique to the nonsense errors acting as a hypothesized source of selection on codon usage.

Cursory support for the hypothesis that nonsense errors help shape codon usage patterns comes from experimental data indicating that many ribosomes fail to complete translation of a given protein (Manley 1978; Jorgensen and Kurland 1990; Kurland 1992). The detection of intragenic patterns of codon bias also provides support for this hypothesis. For example, Hooper and Berg (2000) found evidence for an increase in bias with position in *Escherichia coli*. More recently, Qin et al. (2004) detected clear patterns of increasing codon usage bias with position in *Saccharomyces cerevisiae* and other microorganisms. Similar patterns have been detected indirectly in *S. cerevisiae* as well (Gilchrist and Wagner 2006).

These findings indicate that nonsense errors are costly enough to result in detectable patterns of adaptive codon usage in the sequences of many genes. Given that such patterns of codon usage exist, one is inclined to ask, "What information can be taken from such patterns and how can it be properly interpreted?" This study attempts to answer this question by developing a stochastic model, SEMPPR. Essentially, SEMPPR makes inferences about the production rate of a gene based on its elevation on the fitness landscape of protein production costs.

It should be noted that SEMPPR only considers the effects of synonymous substitutions on a protein's production cost. This is because SEMPPR takes the amino acid sequence of a gene as a given piece of information. Thus, the term "adaptation," as used here, does not refer in anyway to the function of the protein encoded. Instead, here the adaptation of a codon sequence refers the state of its expected cost of producing a protein relative to the minimal possible cost (i.e., the codon sequence, which uses only "optimal" codons).

If nonsense errors are an important source of selection driving the evolution of codon bias, then the strength of this

Key words: codon bias, nonsense errors, fitness landscape, synonymous substitution, evolution.

E-mail: mikeg@utk.edu.

Mol. Biol. Evol. 24(11):2362–2372. 2007
doi:10.1093/molbev/msm169
Advance Access publication August 16, 2007

selection on a gene should be a function of the production rate of the protein it encodes. This selection will be counteracted by mutation and drift. Intuitively, genes observed to be higher on their fitness landscape are more likely to be under stronger selection than genes observed to be lower on their fitness landscape. By employing Bayes' theorem, the work developed here formalizes this intuition in the form of a nested stochastic model SEMPPR. Although the utility of SEMPPR is demonstrated using *S. cerevisiae*, the models and ideas behind its formulation are quite general, making it theoretically applicable to other organisms.

In addition to being able to predict the production rate of a gene from sequence data, SEMPPR should provide a means for calculating the selective force for or against a synonymous substitution. The force is expected to vary as a function of codon position and gene expression level, both of which can be calculated using SEMPPR. Further, because SEMPPR links genotype and fitness in a clear and mechanistic way, the model provides a concrete system for the future examination of more fundamental topics in evolutionary biology, such as the shape and nature of fitness landscapes and behavior of populations on these landscapes.

Materials and Methods

Definitions for all symbols used in this study are listed in Table 1. The presentation of SEMPPR is broken into two parts. The first part focuses on its construction from models of protein production and allele fixation. The second part illustrates how to implement SEMPPR using the 5,847 verified genes of the model organism *S. cerevisiae*. The performance of SEMPPR is then evaluated by comparing its predictions with empirical estimates of protein production rates for individual genes using high-throughput data.

Model Development

In this section, it is shown how the degree to which the allele observed to be fixed in a population reduces the cost of nonsense errors can be used to make inferences about the production rate of the protein it encodes. This is done by mapping genotype to phenotype, phenotype to fitness, and then fitness to fixation probability. As a last step, Bayes' theorem is applied to the fixation probability. The resulting output from SEMPPR is a posterior probability distribution for the protein production rate of a gene based on its observed codon sequence.

From Genotype to Phenotype: Codon Usage and the Cost of Protein Production

When considering the cost of producing a protein in the face of nonsense errors, it is critical to recognize that for every protein which is completely translated, a number of incomplete proteins are also produced. These incomplete proteins are the result of nonsense errors. The cost of these nonsense errors is a function of their expected number and the length of the incomplete proteins. This cost, along with that of the completed protein, must be accounted for when calculating the total cost of protein production.

In the protein translation model presented in Gilchrist and Wagner (2006), a gene is represented as a vector of codons $\vec{c}=\{c_1, c_2, \ldots c_n\}$, where c_i is the elongation rate of the ith codon and n is the number of codons to be translated. Specific values for these parameters can be estimated from sequence and experimental data (see Gilchrist and Wagner 2006 and below).

Based on these assumptions, the probability of a ribosome successfully translating up to and including the ith codon in the sequence $\vec{c}$ is

$$\sigma_i(\vec{c})=\prod_{j=1}^{i}\frac{c_j}{c_j+b}, \tag{1}$$

where b is a constant background nonsense error rate. Although there are data which suggest that b may vary from codon to codon (Freistroffer et al. 2000), for simplicity b is assumed to be constant.

From this definition, it follows that the expected energetic cost of a nonsense error is

$$\xi(\vec{c})=\frac{1}{1-\sigma_n(\vec{c})}\sum_{i=1}^{n}(a_1+a_2\,i)\,\sigma_{i-1}(\vec{c})\frac{b}{c_i+b}. \tag{2}$$

The terms within the summation in equation (2) represent the cost of translating up to the ith codon, the probability a ribosome reaches the ith codon, and the probability a nonsense error occurs at that codon, respectively. The terms a_1 and a_2 represent the cost of ribosome initiation and peptide elongation and are equal to 4 and 2 high-energy phosphate bonds ($\sim P$), respectively. The factor $1/(1-\sigma_n(\vec{c}))$ scales the expected cost in the summation by the probability a nonsense error will occur at some point during the translation process. This factor was neglected in the formulation of ξ originally derived in Gilchrist and Wagner (2006).

The cost function $\xi(\vec{c})$ represents the expected "cost" of a single nonsense error for the given codon sequence $\vec{c}$. This cost function does not take into account the expected "number" of errors that might occur for each complete protein produced. Letting the index i represent the number of failures until a successful completion (which can range from 0 to ∞) and $\eta(\vec{c})$ represent the total expected cost of producing a complete protein from $\vec{c}$ then,

$$\eta(\vec{c})=\sum_{i=0}^{\infty}(1-\sigma_n(\vec{c}))^i\sigma_n(\vec{c})(\xi(\vec{c})i+(a_1+a_2\,n)) \tag{3}$$

$$=\left(\frac{1}{\sigma_n(\vec{c})}-1\right)\xi(\vec{c})+(a_1+a_2\,n). \tag{4}$$

The first 2 terms in the summation of equation (3) represent the probability of i failures occurring before the next successful translation. The last term in the summation represents the expected cost of these i failures and 1 success.

From Phenotype to Fitness: Linking the Energetics of Protein Production to Fitness

The term $\eta(\vec{c})$ represents the expected cost of producing a single, complete protein (protein production cost, for brevity). The rate at which an organism must allocate energy to the production of this protein is the product of this cost and the target production rate of completed proteins ϕ. Presumably, the energy an organism has at its disposal is limited. Consequently, the fitness of an allele will be some function of the energy it requires to produce the protein it encodes. Specifically, SEMPPR assumes that the fitness w of an allele decreases exponentially with the energy expenditure rate it requires, that is,

$$w(\eta)=a_3\exp[-q\,\phi\,\eta], \tag{5}$$

where q is a scaling term and a_3 is a proportionality constant whose value will cancel out in the next calculation.

Unlike a linear assumption, the assumption of a decreasing exponential fitness function ensures that the fitness of an allele never drops below 0. This assumption is also consistent with those made in the population genetics model employed below. This assumption is less restrictive than one might assume given that when $q \times \phi \times \eta \ll 1$, the decrease in $w(\eta)$ with η is approximately linear.

From Fitness to Fixation: Calculating the Probability of Observing an Allele

The next step in formulating SEMPPR is linking fitness and fixation probability of an allele with its given codon sequence. Recently, a number of different researchers have independently derived the stationary distribution for gene fixation (Berg and Lässig 2003; Berg et al. 2004; Sella and Hirsh 2005). Using the formulation presented by Sella and Hirsh (2005), for a haploid organism with a large effective population size, $N_e \gg 1$, the probability a specific allele with fitness w is fixed at a given locus is approximately,

$$f(w)=\frac{w^{N_e}}{\int_0^1 g(w)w^{N_e}\,\mathrm{d}w}, \tag{6}$$

where $g(w)$ represents the probability density function of the set of all possible alleles for a given locus. For reasons of simplicity, it is assumed the mutation rates between different genotypes are symmetric; this assumption can be relaxed in future studies (Sella and Hirsh 2005).

Using the above definitions, the fixation probability of allele in equation (6) can be reformulated on a per locus basis. That is, the fixation probability of a particular allele i with a codon sequence $\vec{c}_i$ and a corresponding protein production cost $\eta_i=\eta(\vec{c}_i)$ is

$$f(\eta_i|\phi)=\frac{\exp[-N_e q\phi\eta_i]}{|\vec{c}|\int_{\eta_{\min}}^{\eta_{\max}} g(\eta')\exp[-N_e q\phi\eta']\mathrm{d}\eta'}, \tag{7}$$

where the parameters q and N_e are implicitly given and $g(\eta')$ represents the probability density function of protein production costs for the focal gene across its synonymous sequence space (i.e., the number of different codon sequences that can produce the same amino acid sequence). The term $|\vec{c}|$ represents the size of this space. This size is equal to the product of the number of codon synonymous at each amino acid position along an entire sequence. Using *S. cerevisiae*'s genome as an example, these synonymous space sequences range in magnitude from 10^8 for very short genes to $>10^{2300}$ for very long genes. The terms $\eta_{\min}$ and $\eta_{\max}$ represent the minimum and maximum protein production costs for the amino acid sequence corresponding to the focal locus. Conceptually, these values correspond to the highest peak and lowest valley of a gene's fitness landscape. Although η values are actually discrete, the set of their sizes is extremely large relative to their range, and the difference between the ith and the $i+1$th highest values is presumably very small. Therefore, for simplicity the distribution of η values are approximated in equation (7) with a continuous distribution.

From Fixation to Production: Inferring Production from Adaptation via Bayes' Theorem

Equation (7) calculates the probability that an allele with a particular codon sequence becomes fixed at a particular locus conditional on the protein production rate ϕ and other information such as q and N_e. In contrast, SEMPPR calculates the posterior probability density function of a production rate given the protein production cost of the observed codon sequence η_{obs}. Therefore, in order to define SEMPPR, Bayes' theorem is applied to equation (7) producing,

$$f(\phi|\eta_{\text{obs}})=\frac{f(\eta_{\text{obs}}|\phi)f(\phi)}{\int_0^{\phi_{\max}} f(\eta_{\text{obs}}|\phi)f(\phi)d\phi}, \tag{8}$$

where η_{obs} is the protein production cost of the observed sequence, $f(\phi)$ represents the prior probability of ϕ, and ϕ_{max} represents the maximum possible protein production rate. Due to the fact that $f(\eta_{\text{obs}}|\phi)$ does not conform to any standard distribution, the proper conjugate prior distribution, if it exists at all, is unknown. Numerical analyses indicate that reasonable the prior distributions $f(\phi)$ include a flat distribution, a truncated log-normal distribution or a composite log-normal distribution. These analyses also indicate that as long as ϕ_{max} is sufficiently large such that $f(\eta_{\text{obs}}|\phi)$ and $f(\phi)$ are very small and decreasing at ϕ_{max}, its exact value has no discernible effect on the predictions made by SEMPPR.

Model Application

This second part illustrates SEMPPR's implementation and performance using the model organism *S. cerevisiae*. Implementing SEMPPR, involved a number of distinct steps. First, the model of protein translation upon which SEMPPR is based was parameterized. Second, the observed protein production cost η was directly calculated for each gene and the distribution of its expected protein production costs $g(\eta)$ was estimated by randomly sampling from the distribution. Third, empirical measurements summarized in Beyer et al. (2004) were used to calculate a set of observed protein production rates. The above results, along with a proxy for N_e, were used to estimate the fitness function's scaling factor q. With this done, summary statistics of the posterior distributions for protein production $f(\phi|\eta_{\text{obs}})$ produced by SEMPPR were compared with their corresponding empirical estimates. To provide a context for these results, the relationship between various codon bias indices to the same empirical data were also evaluated. In addition, internal evaluations of the reliability of the empirical data were also conducted.

On a technical note, initial development and evaluation of SEMPPR was carried out using Mathematica 5.2 (Wolfram Research Inc. 2005). For computational efficiency and portability, these routines were then reimplemented on a Linux workstation in PERL and C using standard libraries. This final version of the code is available in the supplementary data (Supplementary Material online). Updates and bug fixes can be found at www.tiem.utk.edu/∼mikeg/software.

Translation Model Parameterization

The model of protein translation was parameterized as in Gilchrist and Wagner (2006). Briefly, information in Ikemura (1985), Percudani et al. (1997), and Akashi (2003) was used to estimate of the concentration of each tRNA species in *S. cerevisiae*. These values were scaled by a proportionality constant so that the average elongation rate across all codons was 10 amino acids/s after adjusting for wobble (Thomas et al. 1988; Curran and Yarus 1989; Kruger et al. 1998). The nonsense error rate was set to $b = 0.00515$/s after Jorgensen and Kurland (1990) and Tsung et al. (1989).

Coding and Amino Acid Sequences

The 12 May 2006 release of the *S. cerevisiae* genome was downloaded from Saccharomyces genome database (SGD) (Dolinski et al. 2006). Analyses were restricted to the 5,847 verified nuclear genes that lack internal stops. The codon sequence presented in the SGD database was assumed to be fixed in the population and used to calculate η_{obs} for each locus. The amino acid sequence for each gene was used to estimate its distribution of protein production costs $g(\eta)$ as described below.

Simulating the Production Cost Distribution

Given the enormity of the synonymous sequence spaces, the distribution of η values for a given gene, $g(\eta)$, was estimated through simulation. For each gene, 10,000 simulated codon sequences were constructed by randomly selecting one of the possible synonymous codons for each amino acid in the sequence. The protein production cost η for each randomly constructed sequence was then calculated.

Despite the fact that $g(\eta)$ could not be estimated directly, calculation of the upper and lower bounds for each locus was still possible. Specifically, the minimal protein production cost η_{min} is achieved when a codon sequence contains only the fastest translating codons for the given amino acid sequence of a gene. Similarly, the maximal protein production cost η_{max} is achieved when a codon sequence contains only the slowest translating codons.

The cost distribution $g(\eta)$ for each gene was approximated by rescaling a beta distribution as,

$$g(\eta)=\frac{\Gamma(\alpha+\beta)}{\Delta\eta\Gamma(\alpha)\Gamma(\beta)}\left(\frac{\eta_{\mathrm{max}}-\eta}{\Delta\eta}\right)^{\beta-1}\left(\frac{\eta-\eta_{\mathrm{min}}}{\Delta\eta}\right)^{\alpha-1}, \quad (9)$$

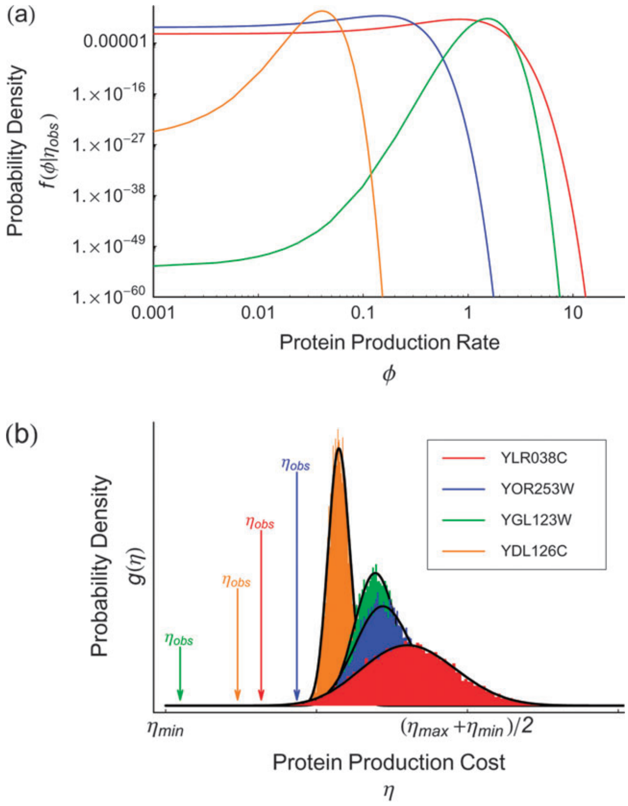

FIG. 1.—(*a*) Probability distributions of protein production costs η for 4 focal genes: YLR038C, YOR253W, YGL123W, and YDL126C. Distributions were estimated by fitting a rescaled beta distribution to a set of 10,000 randomly created synonymous sequences. In order to facilitate their presentation, each gene's η values are presented relative to their individual η_{min} and η_{max} values. The protein production cost values for the sequence observed in SGD are indicated as η_{obs}. (*b*) Posterior probability distributions of the protein production rate ϕ for the same 4 genes. Note that ordering the genes by their observed costs η_{obs} does not necessarily correspond to their ordering with regard to their modal value and that the width of the curves varies between genes.

where $\Delta\eta = \eta_{\mathrm{max}} - \eta_{\mathrm{min}}$ and Γ represents the gamma function. The parameters of rescaled beta distribution, α and β, were estimated using a Bayesian approach with a flat, improper prior of $f(\alpha) = 1$ and $f(\beta) = 1$. Despite the use of an improper prior, the posterior distributions of α and β are proper, and the posterior modes were used for the estimates of the parameters. The values for the posterior modal values of α and β for each gene can be found in supplementary table S1 (Supplementary Material online).

Using a rescaled beta distribution for $g(\eta)$ results in the denominator of equation (7) taking the form of a confluent hypergeometric function: ${}_1F_1(\alpha;\ \alpha + \beta,\ \Delta\eta z)$, where $z = N_e \times q \times \phi \times \eta_{\mathrm{obs}}$. Evaluating ${}_1F_1$ can be numerically intensive, especially when $\Delta\eta z$ is large. However, Butler and Wood (2002) provide a Laplace approximation of this function, which was used in this work. This approximation results in 6th order accuracy but is approximately 2.5 orders faster to calculate.

To illustrate the sampling of $g(\eta)$ and the fitting of a rescaled beta distribution to this data, the simulation results, along with the observed protein production costs η_{obs}, are presented in figure 1*a* for 4 focal genes: YLR038C, YOR253W, YGL123W, and YDL126C. These genes were arbitrarily chosen to cover a wide range of the observed values of (α, β) (supplementary fig. S1*a*, Supplementary Material online) and are involved in distinctly different biological functions. For example, YLR038C is part of the mitochondrial electron transport chain, and YDL126C is involved in the retrotransportation of ubiquinated proteins from the endoplasmic reticulum to the cytosol. As figure 1*a* illustrates, the rescaled beta distribution is able to capture the shape of the observed distribution of the simulated population of η very well.

The overall complexity of the posterior distribution in equations (7) and (8) precludes most direct analysis. However, at least one analytic insight can be made. By examining the slope of $f(\phi|\eta_i, g(\eta), q, N_e)$ at $\phi=0$ and under the assumption of a flat prior for $f(\phi)$, it follows that the posterior mode of ϕ is at the 0 boundary when the protein production cost of the wild-type allele is greater than the average value across genotype space, that is $\eta_{\text{obs}} > \bar{\eta} = (\alpha\eta_{\max} + \beta\eta_{\min})/(\alpha+\beta)$. Work with more complex priors suggests that this result is quite general.

Empirical Estimates of Protein Production Rates

Beyer et al. (2004) provide estimates of the protein production rates for much of the *S. cerevisiae* genome. Conceptually, protein production rates can be broken into 2 components (the translation rate per mRNA and mRNA abundance, τ and m, respectively), such that $\phi = \tau \times m$. The estimates of τ and m for individual genes are obtained from a combination of mRNA ribosome occupancy and abundance measurements presented in Beyer et al. (2004).

The estimates of the translation rate per mRNA τ are derived from mRNA ribosome occupancy measurements. These measurements come from 2 separate studies by Arava et al. (2003) and MacKay et al. (2004). Because the error in the occupancy data is poorly understood, Beyer et al. (2004) simply averaged these measures and then multiplied by 10 amino acids/sec to generate a translation rate per mRNA. It should be noted that this method of estimating τ ignores any impact heterogeneity in codon elongation rates or nonsense errors may have on ribosome occupancy. Given that the occupancy measurements are dependent on florescence and migration through a density gradient, the error in these measurements is likely to be log-normally distributed. As a result, the geometric mean, as opposed to the arithmetic mean, of the observations was used as the estimate of τ, that is, $\tau = \sqrt{\tau_1 \tau_2}$. Using the arithmetic mean has no substantial effect on the evaluation of SEMPPR's performance.

The estimates of mRNA abundance m are based on Affymetrix GeneChips measurements from multiple labs. The authors modified some of the measurements to counteract the saturation effect observed in high-expression genes (see Beyer et al. 2004 for more details). Due to the large variation in mRNA abundance measurements between experiments, Beyer et al. (2004) propose using the median values as the best estimate of mRNA abundance. Examination of the error in the mRNA abundances reveals that it is also log-normally, rather than normally, distributed. This result reinforces the wisdom of this choice (supplementary figure, Supplementary Material online). This is because for a log-normally distributed random variable M where $M \sim \text{log-N}(\mu, \sigma)$, the central parameter μ and median of the distribution are equivalent. Thus, assuming log-normal measurement error with a flat prior uninformative prior on $ln(m)$ is consistent with previous analyses and allows the generation of posterior probability intervals (PIs) for the mRNA abundance of each gene. Consequently, the posterior mode for M was used as the estimate of mRNA abundance for an individual gene, that is, $\hat{m} = (\prod^n m_i)^{1/n}$. Thus, the empirical estimate of a gene's protein production rate was $\hat{\Phi} = \hat{m} \times \tau$.

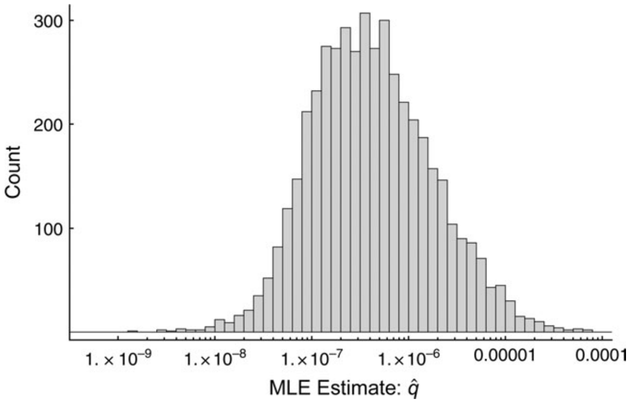

FIG. 2.—Individual MLEs of the scaling coefficient q from 4,633 *Saccharomyces cerevisiae* genes. MLE values were the ones, which maximized the likelihood of the empirically estimated protein production rate $\phi = \hat{\phi}$ given the observed protein production cost η_{obs} using equation (8).

Estimating q from Multiple Genes

The value of q was independently estimated for the 4,633 *S. cerevisiae* genes, with empirical estimates of the protein production rate $\hat{\Phi}$ and the mode of SEMPPR's prediction both greater than 0. Maximum likelihood estimates (MLEs) of q was calculated using equation (7) and by assuming $\phi = \hat{\Phi}$ and using the effective population size of *Saccharomyces paradoxus* $N_e = 1.36 \times 10^7$ as a proxy for *S. cerevisiae* (Wagner 2005). Assuming that the error in the estimates of mRNA and ribosome occupancy are both log-normally distributed, it follows that the empirically estimated protein production rate $\hat{\Phi}$ for an individual gene will be lognormally distributed around its true value and, consequently, the estimates of q should also be log-normally distributed around its actually value. As expected, the distribution of these MLE estimates of q appears to be log-normally distributed around the geometric mean of $\hat{q} = 4.19 \times 10^{-7}$ (fig. 2), which was also the value used to parameterize SEMPPR. This set of independent estimates of q was used to calculate the upper and lower 95% confidence interval (CI) for q as 3.39×10^{-8} and 8.49×10^{-6}, respectively.

Given the 200-fold range in the CI for q and the use of N_e based on *S. paradoxus*, it is worth noting how these terms affect SEMPPR's predictions. In general, increasing values of q or N_e is equivalent to increasing the importance of selection relative to drift on allele fixation probabilities either through stronger selection, q, or a larger population size, N_e. Given the fact that q, N_e, and ϕ are multiplied by one another in equation (7), for a given observed protein production cost η_{obs} increasing either q or N_e by some factor x rescales the probability distribution function for ϕ produced by SEMPPR by $1/x$. Indeed, the difference between N_e for *S. cerevisiae* and *S. paradoxus* is ultimately absorbed in the estimates of q. Thus, the 1 unknown parameter in this analysis, q, and 1 questionable parameter, N_e, work together as a scaling factor across all of SEMPPR's predictions (supplementary fig. S4, Supplementary Material online). Because SEMPPR's predictions are evaluated on a log–log scale, changing either N_e or q has no effect on the correlation coefficient between the predicted and observed values.

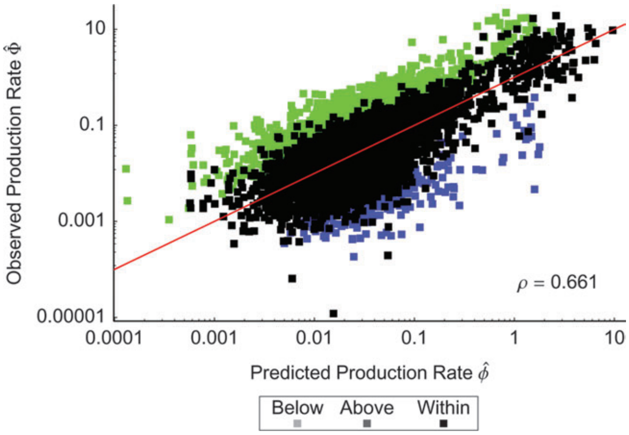

FIG. 3.—Scatter plot of empirical estimates of a protein's production rates $\hat{\Phi}$ versus the geometric mean of SEMPPR's predictions $\hat{\phi}$ for 5,432 genes in *Saccharomyces cerevisiae*. Color scheme indicates whether or not the 95% PI of the observed and predicted rates of ϕ overlapped. Green points are where the predicted 95% PI were below the empirically based 95% PI. Black points are where the predicted 95% PI overlapped with the empirically based 95% PI. Blue points are where the predicted 95% PI were above the empirically based 95% PI. The correlation coefficient ρ between the observed and predicted values was 0.661.

Results

Evaluating SEMPPR

Predicting Empirical Estimates of ϕ

Having all of the necessary components in hand, SEMPPR was used to calculate the posterior distributions of the protein production rate, $f(\phi|\eta_{\text{obs}})$, for each of the 5,847 genes analyzed. Figure 1b illustrates SEMPPR's distributions for the 4 focal genes mentioned earlier. The cumulative distribution functions for these and the 5,843 other verified *S. cerevisiae* genes are presented in supplementary table S2 (Supplementary Material online). The results from SEMPPR can be used to calculate any of a number of statistics such as the 95% PI or measures of central tendencies such as the posterior mode, geometric mean, or arithmetic mean. All 3 measures of central tendencies are highly correlated with one another on a log-log scale (i.e., the correlations coefficients ρ range from 0.96 to 0.99). These values are presented in supplementary table S3 (Supplementary Material online). For simplicity, the geometric mean of SEMPPR's posterior distribution was chosen to represent the predicted protein production rate, that is, $\hat{\phi}=\exp(\int_0^{\phi_{\max}} \ln(\phi) f(\phi|\eta_{\text{obs}}\, d\phi))$.

Because there is uncertainty in both the predicted and observed production rates, $\hat{\phi}$ and $\hat{\Phi}$, standard goodness of fit models cannot be used. Instead, the performance of SEMPPR's predictions was evaluated by the correlation between the $\log_{10}(\hat{\phi})$ and $\log_{10}(\hat{\Phi})$, indicating how much of the variation in $\log_{10}(\hat{\Phi})$ is explained by $\log_{10}(\hat{\phi})$. The correlation coefficient ρ between these predicted and the empirical estimates of protein production across the *S. cerevisiae* genome is 0.661 (fig. 3). This high correlation coefficient indicates that SEMPPR's $\log_{10}(\hat{\phi})$ values are highly correlated with the empirically $\log_{10}(\hat{\Phi})$ values.

For comparison, similar correlation analyses were conducted with 4 commonly used indices of codon usage bias: the Codon Adaptation Index (CAI) (Sharp and Li 1987), Codon Bias Index (Bennetzen and Hall 1982), Frequency of Optimal Codons (Ikemura 1981, 1985), and Effective Number of Codons (Wright 1990). The correlation coefficients between these indices and $\log_{10}(\hat{\Phi})$ range from 0.574 to 0.657 (fig. 4). Thus even in its initial formulation, SEMPPR is able to predict the expression level of genes as well as the standard indices.

Overlap between SEMPPR's Predictions and Estimates

In addition to having a higher correlation with the empirical data, SEMPPR has 2 distinct advantages over index-based predictions. The first advantage is that SEMPPR's predictions are in units of a biologically meaningful and measurable process, that is, protein production rate of a gene ϕ. The second advantage is the probabilistic nature of SEMPPR's predictions. The fact that the frequency at which SEMPPR's 95% PI overlaps with empirical estimates of ϕ can be calculated illustrates both of these advantages.

Because $\hat{\Phi}$ values are themselves estimates with relatively large error, 2 separate analyses were conducted. In the first analysis, the error in estimating $\hat{\Phi}$ was ignored and whether or not the SEMPPR's 95% PI overlaps with $\hat{\Phi}$ was determined. Because the first analysis ignores the substantial measurement error in $\hat{\Phi}$, the performance of SEMPPR's predictions will be underestimated. To account for part of the measurement error, a second analysis was conducted where the measurement error in mRNA abundance $\hat{m}$ was used as a proxy for the measurement error in $\hat{\Phi}$.

In the first analysis in which the error in $\hat{\Phi}$ was ignored, SEMPPR's 95% PI for ϕ overlap with $\hat{\Phi}$ for 42.4% of the genes. For 36.7% and 20.9% of the genes, SEMPPR's 95% PI are below and above $\hat{\Phi}$, respectively. Because q acts as a scaling factor on SEMPPR's predictions, changing q will change the range of SEMPPR's 95% PI. Figure 5*a* illustrates how this behavior affects the overlap between SEMPPR's predictions and $\hat{\Phi}$. The overlap between SEMPPR's 95% PI and the observed $\hat{\Phi}$ values peaks at 45% when $q = 1.9 \times 10^{-7}$ and has equal frequencies (27.9%) of above and below errors at $q = 2.96 \times 10^{-7}$.

In the second analysis, the 95% PI of the mRNA measurements m for each gene was calculated based on the Affymetrix GeneChip measurements used in Beyer et al. (2004). The error was assumed to be log-normally distributed with a flat prior on $\log_{10}(m)$. Only genes with more than 17 measurements were included. The 95% PI values were then multiplied by the estimated per mRNA translation rate

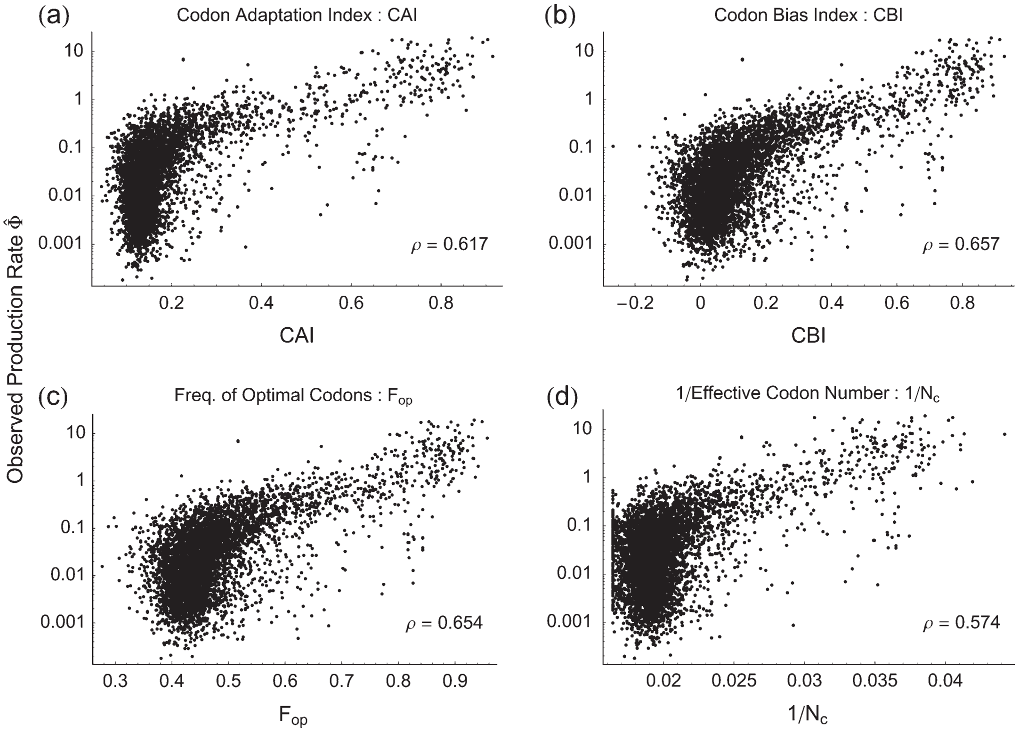

FIG. 4.—Comparison of various indices and the logarithm of the empirical estimates of a protein's production rate $\hat{\Phi}$ including their correlation coefficients ρ. (*a*) CAI (Sharp and Li 1987), (*b*) Codon bias index (Bennetzen and Hall 1982), (*c*), Frequency of optimal codons F_{op} (Ikemura 1981, 1985), and (*d*) the inverse of the effective number of codons $1/N_{\mathrm{c}}$ (Wright 1990). Index values were calculated using the program CodonW (Peden 2005).

$\hat{\tau}$ to generate proxy 95% PI for $\hat{\Phi}$. Given that there is measurement error in the estimates of both m and τ, the PIs for m are underestimates of the uncertainty in $\hat{\Phi}$. In this second analysis, SEMPPR's 95% PI for ϕ overlap with the $\hat{\Phi}$ 95% PI for 76.0% of the genes. For 18.6% and 6.41% of the genes, SEMPPR's 95% PI are below and above $\hat{\Phi}$, respectively. As before, the degree of overlap between SEMPPR's predictions and the empirical data is dependent on the value of the scaling parameter q (fig. 5*b*). The overlap between SEMPPR's and the $\hat{\Phi}$'s 95% PI peaks at 78.8% when $q = 2.35 \times 10^{-7}$ and have equal frequencies (10.9%) of above and below errors at $q = 2.70 \times 10^{-7}$.

Discussion

The idea of linking gene expression and codon bias within a framework of selection, mutation, and drift is not new (e.g., see Bulmer 1991; Kurland 1992; Kliman and Hey 1994; Comeron and Kreitman 2002). What is new in this work is the idea of linking these processes in an explicitly mechanistic manner to develop SEMPPR. SEMPPR begins by linking the codon sequence of a genotype to its phenotype, that is, its protein production cost. The phenotype is then linked to fitness, and the probability of an allele being fixed is calculated using a population genetics model. As a last step Bayes' theorem is used, resulting in SEMPPR's ability to make inferences about a gene's protein production rate based on its observed codon sequence.

Because SEMPPR's is in part built on a model of population genetics, some of the metaphors from that field can be used for a more intuitive understanding of how SEMPPR works. Conceptually, the sequences with the minimal and maximal protein production costs, η_{min} and η_{max}, represent the highest peak and lowest point of a gene's protein production cost fitness landscape. The location of the observed sequence's cost η_{obs} on this landscape reflects the combined processes of selection, mutation, and drift. The height of the peaks and valleys of the fitness landscape are scaled by production rate of the gene ϕ. As a result, the difference in energetic usage between the highest peak and lowest valley of the fitness landscapes is small for genes with low production rates and large for genes with high production rates.

Intuitively, proteins with higher production rates are expected to show greater degrees of adaptation due to the increased importance of selection. Although this is essentially what is observed, equation (8) indicates that SEMPPR's inferences about the production rate of a gene is not just a function of the absolute or relative distance between the observed η_{obs} and the minimal η_{min} protein production costs. Instead, SEMPPR's inferences are based on where η_{obs} lies with respect to the entire set of possible protein production costs $g(\eta)$ (fig. 1 and supplementary fig. S3 [Supplementary Material online]).

The strong performance of SEMPPR suggests that the one analytical result presented here, where the mode posterior production rate is 0 when an observed allele's protein production cost is less than the expected value from $g(\eta)$, provides a conceptually clear and useful criteria for identifying pseudogenes (which by definition have 0 expression).

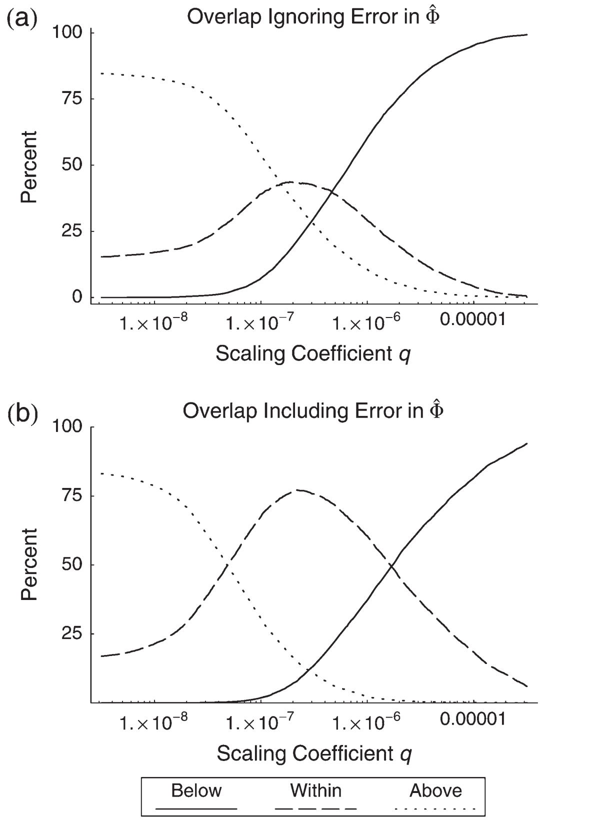

FIG. 5.—Illustration of how changing the scaling coefficient q affects the degree of overlap between the empirically observed and predicted production rates, $\hat{\Phi}$ and $\hat{\phi}$, respectively. (*a*) Assumes that the empirical measurements are without any error, that is, $\phi = \hat{\Phi}$. (*b*) Uses estimated error in mRNA abundance measurements $\hat{m}$ as a proxy for the error in $\hat{\Phi}$. This error is actually a lower bound for the error in $\hat{\Phi}$ because it does not consider the error introduced by translation rate per mRNA $\hat{\tau}$.

More generally, the process of linking genotype to phenotype and phenotype to fitness developed here provides an explicit framework for studying the structure and evolutionary behavior on a biologically realistic fitness landscape. Similar studies on fitness landscapes for transcription factor binding (Berg et al. 2004) and protein thermostability (Bastolla et al. 2006) have already produced interesting results, suggesting that the study of fitness landscapes on a molecular scale will become an increasingly important area of research.

One result from the development of SEMPPR is that, from the perspective of adaptation with respect to nonsense errors, codon bias should be a function of a gene's protein production rate. This idea is in contrast to previous studies, which have tried to link codon bias to the mRNA abundance of a gene (which affects its net protein production rate) or its protein abundance (which is affected by its net protein production rate). Recognizing that it makes biological sense to link codon usage to protein production rate rather than just mRNA or protein abundance is an important conceptual advantage of using mechanistic models. Despite their importance in the identification and initial description of codon bias, it is quality that cannot be obtained using heuristically derived metrics.

Although the population genetics model SEMPPR use requires an exponential relationship between trait values and fitness, because the scaling coefficient q appears to be very small, one could actually approximate the exponential fitness function as a linear function using a Taylor Series. Doing so yields the selection coefficient used in traditional population genetics of approximately $s = -q(\eta - \eta_{\mathrm{min}})\phi$. For most observed alleles this translates into selection coefficients relative to the optimal allele ranging from 10^{-6} to 10^{-4}. Given the large effective population size of *S. cerevisiae*, if one were simply considering competition between 2 alleles, then one would expect the optimal allele to eventually emerge and go to fixation because $sN_e \gg 1$. However, because there are a large number of differences between the observed allele and the optimal allele, it is impossible to mutate directly from one to another. Instead, the relevant selection coefficients are between alleles that are one mutational step from one another. These coefficients are likely to be much smaller than previously published estimates in *E. coli* (Hartl et al. 1994) or *Drosophila* (Akashi 1995).

Although small in size, because codon bias exists, these selection coefficients are clearly still evolutionarily significant. Further, because the substitution of suboptimal codons at the $5'$ end of a sequence has a greater negative effect on fitness than a similar substitution in the $3'$ end, the selection coefficients between the nearest mutational neighbors of an allele are not equal.

The size and nature of the selection coefficients against synonymous substitutions are clearly very subtle. However, these coefficients are important to consider when building phylogenetic models of nucleotide substitution. The framework developed here should provide a means of calculating selection coefficients and, thereby, making it possible to take such effects into account.

Currently, new techniques are being developed to account for such selection against synonymous substitutions by weighting the importance of a substitution based on codon position or specific codon type (Yang and Nielsen 2000; Bierne and Eyre-Walker 2003; Hirsh et al. 2005). However, these techniques are observational, rather than deductive, in nature. Further, none of these techniques take into account the effect of both expression level and codon position. By revisiting the models underlying SEMPPR, it should be possible to calculate the fixation probability of a particular codon given its position and its gene's estimated production rate. These substitution probabilities could then be incorporated into models used for phylogenetic reconstructions. These values could also be incorporated into more recent approaches testing for positive selection (e.g., Zhang et al. 2006).

In terms of the predictive performance of SEMPPR, the protein production rates it predicts are well correlated with the available empirical data ($\rho = 0.661$ on a log–log scale). This strong correlation is in spite of the fact that the uncertainty of the empirical estimates of the protein production rate of a gene is quite large. For example, variation in the mRNA abundance measurements from Affymetrix GeneChips often ranges over 3 orders of magnitude between trials. The ribosome occupancy data upon which the per mRNA translation rates are based also appear to

Table 1
List of Symbols Used

c_i	=	Elongation rate of codon i
$\vec{c}$	=	Set of c_i values for an entire protein
$\lvert\vec{c}\rvert$	=	Size of the synonymous sequence space for a given gene, that is, the number of different ways a protein can be constructed using different combinations of synonymous codons
b	=	Nonsense error rate
$\sigma_i(\vec{c})$	=	Probability a ribosome will translate up to and including the ith codon of the sequence $\vec{c}$
$\sigma_n(\vec{c})$	=	Probability a ribosome will successfully translate all n codons in the sequence $\vec{c}$
$\xi(\vec{c})$	=	Expected cost of nonsense errors given sequence $\vec{c}$
$\eta(\vec{c})$	=	Expected cost of producing a single, complete protein given sequence $\vec{c}$
η_{min}	=	Minimum protein production cost achieved when $\vec{c}$ is composed of only the fastest translating codons
η_{max}	=	Maximum protein production cost achieved when $\vec{c}$ is composed of only the slowest translating codons
η_{obs}	=	Protein production cost for codon sequence observed in the SGD database
$g(\eta)$	=	PDF of the protein production costs η for a single gene
ϕ	=	Target production rate of a given protein
$f(\phi)$	=	Prior probability for the protein production rate of a gene
$f(\eta\lvert\phi)$	=	Probability an allele with protein production cost η is fixed in the population given a protein production rate ϕ
$f(\phi\lvert\eta)$	=	Posterior probability of protein production rate ϕ given allele with protein production cost η is fixed in the population
$w(\eta)$	=	Fitness of allele with protein production cost η
q	=	Scaling coefficient for linking protein production cost η to fitness w
N_e	=	Effective population size
$\hat{q}$	=	Estimate of scaling coefficient q based on the geometric mean of 4,728 individual maximum likelihood estimates
$\hat{\Phi}$	=	Empirical estimate of protein production rate ϕ using data complied in Beyer et al. (2004)
	=	$\hat{m} \times \hat{\tau}$
$\hat{m}$	=	Empirical estimate of mRNA abundances based on the geometric mean of the observed values
$\hat{\tau}$	=	Empirical estimate of the protein translation rate per mRNA based on the geometric mean of the observed values

be quite noisy. For example, the correlation coefficient between the log-transformed data from Arava et al. (2003) and MacKay et al. (2004) have a correlation coefficient of $\rho = 0.55$. Thus, it appears that the quality of SEMPPR's predictions is of similar or higher quality than those based on high-throughput data set. Further, as noted previously, the standard method by which the per mRNA translation rate τ is inferred from ribosomal occupancy data ignores the subtle effects of codon heterogeneity and nonsense errors. Thus taking these effects into account when estimating τ should improve the correlation between SEMPPR and the observed protein production rates.

The predictive performance of SEMPPR is similar to the that of the standard codon bias indices. However, SEMPPR has two important advantages. First is that its output is more interpretable from a biological and statistical perspective. The second is that the assumptions underlying models upon which it is built can always be refined further. For example, in the current formulation SEMPPR assumes that only fully translated proteins are functional. The assumption that only complete proteins have functionality is likely to be overly restrictive because most proteins are likely to maintain most or all of their functionality if all but the last few amino acids are translated. Indirect evidence for this idea is found in the fact that codon bias appears to decrease at the end of a coding sequence (Qin et al. 2004). In contrast, preliminary work with SEMPPR where the final 25 or 50 codons are ignored results in only minor changes to its output. Nevertheless, a more sophisticated analysis might reveal information on the general relationship between protein functionality and completeness.

One likely way SEMPPR could be improved is by refining the protein translation and population genetics models that underlie it. For example, the protein translation model assumes that ribosomes moving along a transcript do not interfere with one another. This assumption likely holds for most eukaryotes but is unlikely to hold for many prokaryotes. Works by Chou (2003) and Basu and Chowdhury (2007) suggest ways by which inter-ribosomal interference could be incorporated into the translation model used here. The translation model could also be further refined by allowing slight variations in the background nonsense error rates between different codons. Similarly, the population genetics model could be further refined to incorporate nonuniform mutation rates between codons.

One significant shortcoming of SEMPPR that does not have such a clear solution is its implicit assumption of evolutionary independence between codons and loci. This assumption ignores effects from forces such as interference selection, which also appears to play a role in the evolution of codon bias (Comeron and Kreitman 2002).

Ideally, in the future it will be possible to use SEMPPR to make quantitative predictions about protein production rates from the genomes of completely sequenced nonmodel organisms. This is because applying SEMPPR across a genome only requires knowledge of a relatively small number of parameters such as codon elongation rates and the background nonsense error rate of an organism. Currently, tRNA copy number can be accurately estimated from genomic data (Lowe and Eddy 1997; Hallin and Ussery 2004), and Akashi (2001) shows that elongation rates for different codons can be inferred from such data. In addition, work with CAI by Carbone et al. (2003) suggests that the codon elongation rates could be estimated from sequence data alone. Alternatively, information on these parameters from both tRNA copy number and sequence data could be combined using a Bayesian approach. Given the hundreds of nonmodel microbial organisms now being sequenced, SEMPPR could be enormously useful for understanding the ecological roles played by these less studied but nonetheless ecologically important species. This use of SEMPPR is similar in spirit to the work by Carbone et al. (2005), but SEMPPR's predictions would be more mechanistically based and promise to be easier to interpret.

Supplementary Material

Supplementary data, figures S1–S4, and tables S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Addendum

The index values calculated by CodonW in the original Advanced Access version of this manuscript posted on August 16, 2007 had a significant number of errors that appear to be due to the long ORF description fields found in SGD's *orf_coding. fasta* file. These errors caused all of the indices to appear to perform substantially worse than they actually do. This problem was corrected on September 6, 2007.

Acknowledgments

I would like to thank Aaron King, Jeffery Thorne, Lou Gross, Igor Zhulin, Russell Zaretzki, and Andreas Wagner for their helpful discussions with this work. I would also like to thank the 2 anonymous reviewers for their constructive and encouraging criticisms and A. Beyer for providing me with his database of Affymetrix measurements. In addition I am very grateful to Morten Kloster for pointing out the errors in the codon index values used in the original Advanced Access version of this manuscript. Premal Shah generated the PERL and C code. Financial support was provided by the University of Tennessee, Knoxville.

Literature Cited

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics. 136:927–935.

Akashi H. 1995. Inferring weak selection from patterns of polymorphism and divergence at silent sites in *Drosophila* DNA. Genetics. 139:1067–1076.

Akashi H. 2001. Gene expression and molecular evolution. Curr Opin Genet Dev. 11:660–666.

Akashi H. 2003. Translational selection and yeast proteome evolution. Genetics. 164:1291–1303.

Akashi H, Eyre-Walker A. 1998. Translational selection and molecular evolution. Curr Opin Genet Dev. 8:688–693.

Arava Y, Wang YL, Storey JD, Liu CL, Brown PO, Herschlag D. 2003. Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA. 100:3889–3894.

Bastolla U, Porto M, Roman HE, Vendruscolo M. 2006. A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the protein data bank. BMC Evol Biol. 6. doi:10.1186/1471-2148-6-43.

Basu A, Chowdhury D. 2007. Traffic of interacting ribosomes: effects of single-machine mechanochemistry on protein synthesis. Phys Rev E. 75. doi:10.1103/PhysRevE-75.021902.

Bennetzen JL, Hall BD. 1982. Codon selection in yeast. J Biol Chem. 257:3026–3031.

Berg J, Lässig M. 2003. Stochastic evolution and transcription factor binding sites. Biophysics. 48:S36–S44.

Berg J, Willmann S, Lassig M. 2004. Adaptive evolution of transcription factor binding sites. BMC Evol Biol. 4. Art. No. 021902 Part 1. doi:10.1186/1471-2148-4-42.

Beyer A, Hollunder J, Nasheuer HP, Wilhelm T. 2004. Post-transcriptional expression regulation in the yeast *Saccharomyces cerevisiae* on a genomic scale. Mol Cell Proteomics. 3:1083–1092.

Bierne N, Eyre-Walker A. 2003. The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. Genetics. 165:1587–1597.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics. 129:897–907.

Butler RW, Wood ATA. 2002. Laplace approximations for hypergeometric functions with matrix argument. Ann Stat. 30:1155–1177.

Carbone A, Kepes F, Zinovyev A. 2005. Codon bias signatures, organization of microorganisms in codon space, and lifestyle. Mol Biol Evol. 22:547–561.

Carbone A, Zinovyev A, Kepes F. 2003. Codon adaptation index as a measure of dominating codon bias. Bioinformatics. 19:2005–2015.

Chou T. 2003. Ribosome recycling, diffusion, and mRNA loop formation in translational regulation. Biophys J. 85:755–773.

Comeron JM, Kreitman M. 2002. Population, evolutionary and genomic consequences of interference selection. Genetics. 161:389–410.

Curran JF, Yarus M. 1989. Rates of aminoacyl-trans-RNA selection at 29 sense codons invivo. J Mol Biol. 209: 65–77.

Dolinski K, Balakrishnan R, Christie KR, et al. 2006. (20 coauthors). 2006. Saccharomyces genome database [Internet]. [released 2006 May 12]. Available from: ftp://ftp.yeastgenome.org/yeast/.

Elf J, Nilsson D, Tenson T, Ehrenberg M. 2003. Selective charging of tRNA isoacceptors explains patterns of codon usage. Science. 300:1718–1722.

Eyre-Walker A. 1996. Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? Mol Biol Evol. 13:864–872.

Freistroffer DV, Kwiatkowski M, Buckingham RH, Ehrenberg M. 2000. The accuracy of codon recognition by polypeptide release factors. Proc Natl Acad Sci USA. 97:2046–2051.

Gilchrist MA, Wagner A. 2006. A model of protein translation including codon bias, nonsense errors, and ribosome recycling. J Theor Biol. 239:417–434.

Gouy M, Gautier C. 1982. Codon usage in bacteria—correlation with gene expressivity. Nucleic Acids Res. 10:7055–7074.

Greenbaum D, Colangelo C, Williams K, Gerstein M. 2003. Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol. 4:117–124.

Gygi SP, Rochon Y, Franza BR, Aebersold R. 1999. Correlation between protein and mRNA abundance in yeast. Mol Cell Biol. 19:1720–1730.

Hallin PF, Ussery DW. 2004. CBS genome atlas database: a dynamic storage for bioinformatic results and sequence data. Bioinformatics. 20:3682–3686.

Hartl DL, Moriyama EN, Sawyer SA. 1994. Selection intensity for codon bias. Genetics. 138:227–234.

Hirsh AE, Fraser HB, Wall DP. 2005. Adjusting for selection on synonymous sites in estimates of evolutionary distance. Mol Biol Evol. 22:174–177.

Hooper SD, Berg OG. 2000. Gradients in nucleotide and codon usage along *Escherichia coli* genes. Nucleic Acids Res. 28:3517–3523.

Ikemura T. 1981. Correlation between the abundance of *Escherichia-coli* transfer-RNAs and the occurrence of the respective codons in its protein genes—a proposal for a synonymous codon choice that is optimal for the *Escherichia-coli* translational system. J Mol Biol. 151:389–409.

Ikemura T. 1985. Codon usage and transfer-RNA content in unicellular and multicellular organisms. Mol Biol Evol. 2:13–34.

Jorgensen F, Kurland CG. 1990. Processivity errors of gene-expression in *Escherichia coli*. J Mol Biol. 215:511–521.

Kliman RM, Hey J. 1994. The effects of mutation and natural-selection on codon bias in the genes of *Drosophila*. Genetics. 137:1049–1056.

Kliman RM, Irving N, Santiago M. 2003. Selection conflicts, gene expression, and codon usage trends in yeast. J Mol Evol. 57:98–109.

Kruger MK, Pedersen S, Hagervall TG, Sorensen MA. 1998. The modification of the wobble base of tRNA(glu) modulates the translation rate of glutamic acid codons in vivo. J Mol Biol. 284:621–631.

Kurland CG. 1992. Translational accuracy and the fitness of bacteria. Annu Rev Genet. 26:29–50.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25:955–964.

MacKay VL, Li XH, Flory MR, et al. (12 co-authors). 2004. Gene expression analyzed by high-resolution state array analysis and quantitative proteomics–response of yeast to mating pheromone. Mol Cell Proteomics. 3:478–489.

Manley JL. 1978. Synthesis and degradation of termination and premature-termination fragments of beta-galactosidase *in vitro* and *in vivo*. J Mol Biol. 125:407–432.

Peden J. 2005. Codonw. v.1.4.2. Available from: http://sourceforge.net/projects/codonw/ (accessed on 2007 Sept 1).

Percudani R, Pavesi A, Ottonello S. 1997. Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. J Mol Biol. 268:322–330.

Plotkin JB, Dushoff J, Desai MM, Fraser HB. 2006. Codon usage and selection on proteins. J Mol Evol. 63:635–653.

Qin H, Wu WB, Comeron JM, Kreitman M, Li WH. 2004. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. Genetics. 168:2245–2260.

Sella G, Hirsh AE. 2005. The application of statistical physics to evolutionary biology. Proc Natl Acad Sci USA. 102: 9541–9546.

Sharp PM, Li WH. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.

Thomas LK, Dix DB, Thompson RC. 1988. Codon choice and gene-expression—synonymous codons differ in their ability to direct aminoacylated-transfer RNA-binding to ribosomes invitro. Proc Natl Acad Sci USA. 85:4242–4246.

Tsung K, Inouye S, Inouye M. 1989. Factors affecting the efficiency of protein-synthesis in *Escherichia coli*—production of a polypeptide of more than 6000 amino-acid residues. J Biol Chem. 264:4428–4433.

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

Wolfram Research Inc. 2005. Mathematica. version 5.2. Champaign, (IL): Wolfram Research Inc.

Wright F. 1990. The effective number of codons used in a gene. Gene. 87:23–29.

Yang ZH, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17:32–43.

Zhang Z, Li J, Yu J. 2006. Computing k_a and k_s with a consideration of unequal transitional substitutions. BMC Evol Biol Art. 6.

Lauren McIntyre, Associate Editor

Accepted August 1, 2007