Pursuit - The Journal of Undergraduate Research at The University of Tennessee

Volume 4
Issue 2 *Spring 2013*

Article 5

March 2013

Improving Codon Evolution Models Using Complex Mutation Models

Preston Hewgley
University of Tennessee, Knoxville, whewgley@utk.edu

Pursuit: The Journal of Undergraduate Research at the University of Tennessee
Copyright © The University of Tennessee

Pursuit

Improving Codon Evolution Models Using Complex Mutation Models

PRESTON HEWGLEY
Advisor: Michael A. Gilchrist

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

This paper discusses an improvement in a Stochastic Evolutionary Model of Protein Production Rate (SEMPPR) by revising the method by which it models mutation. SEMPPR previously assumed unbiased mutation, an assumption whose inaccuracy is made clear by observed codon counts of low-expression genes, where mutation determines equilibrium state. This paper presents a new, more complex model generalized on a per-codon basis and calculated from observed codon frequencies using a maximum likelihood framework. Results obtained from SEMPPR using the codon specific mutation model proved more accurate in predicting a protein's production rate, reaffirming that complex mechanisms govern codon mutation rates.

Introduction

The presence of more translating codons than translated amino acids in the genetic code creates redundancy in the protein code. Codons that translate the same amino acid are referred to as synonymous codons, and codon sequences that produce the same protein sequence are referred to as synonymous sequences. Codon usage bias, or codon bias, refers to the widely documented phenomenon showing the non-uniform use of synonymous codons within an organism's genome (Ikemura 1981; Bennetzen and Hall 1982; Sharp and Li 1987). Several explanations of codon bias include mutational bias, intron splicing, recombination and gene conversion, DNA packaging, and selection for increased translational efficiency or accuracy (Bernardi and Bernardi 1986; Bulmer 1988, 1991; Shields et al. 1988; Kliman and Hey 1993,1994; Akashi 1994,2003; Xia 1996,1998; Akashi and Eyre-Walker 1998; Musto et al. 2003; Chen et al. 2004; Chamary and Hurst 2005a,b; Comeron 2006; Lin et al. 2006; Warnecke and Hurst 2007; Drummond and Wilke 2008; Warnecke et al. 2008).

The Stochastic Evolutionary Model of Protein Production Rate (SEMPPR), introduced in Gilchrist (2007), assumed unbiased mutation, where mutation rates among all codons are equal. In low-expression genes, mutation determines the equilibrium state of codon bias because few selective pressures exist to influence translational efficiency. Assuming unbiased mutation, low-expression genes should be composed of equal frequencies of all codons for a given amino acid. As shown in Figure 1, non-uniform codon frequencies in low-expression genes suggest that codon mutation rates between codons are unequal. Several mechanistic models have been proposed to explain mutation bias including transition/transversion bias and A-T bias (Kimura 1980; Hershberg and Petrov 2010).

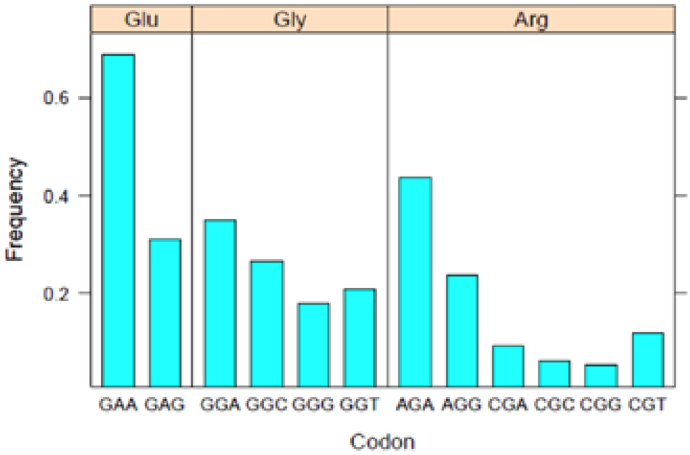

Figure 1. This bar chart shows codon frequencies of low expression genes (< 0.01/sec) in S. Cerevisiae. Note that in unbiased mutation, all codon frequencies should be equal.

The new mutation model presented in this study takes a heuristic approach in which codon-specific mutation rates are calculated using a maximum likelihood framework based on observed codon frequencies (Shah and Gilchrist 2011). The new codon-specific mutation model was implemented in SEMPPR to observe its effect on SEMPPR's ability to predict protein production rates.

Methods

SEMPPR uses an evolutionary framework to predict the production rate of each protein sequence based on the fitness of its codon sequence (Gilchrist 2007). Assuming that a codon sequence is more favorable if it uses less energy to produce a protein, the fitness of a codon sequence is assumed a negative exponential function of the energy expenditure rate for its production, $\phi\eta$.

$$w(\vec{c}) \propto e^{-q\phi\eta}$$

In this equation, ϕ represents the production rate of the protein (units of proteins/second), and η represents the cost of completing one protein (units of energy), and the coefficient q represents a scaling factor, which makes the proportional scale of these units arbitrary. As explained later, protein cost η is calculated as a composition of the cost of one complete polypeptide plus the expected energetic cost due to nonsense errors. The previously mentioned energy expenditure rate $\phi\eta$ is calculated as the rate of protein production ϕ times the cost of producing one protein η.

To conceptualize the idea of this fitness function, consider cases of low-expression genes and high-expression genes. Given the previously defined fitness function, as ϕ approaches 0, w approaches 1 independently of protein cost η. Since fitness, a measure of reproductive success, does not vary with η, all organisms are equally likely to pass on their genetic material when considering only protein cost's effect on fitness. In this case, the codon that is mutated to most frequently by other codons will appear most frequently. Therefore, mutation is the major contributor to the equilibrium state of codon bias in low-expression genes. In contrast, as ϕ gets large, w varies significantly with values of η. A large difference in w is more likely to affect reproductive success, and organisms with low-cost sequences are more likely to pass on their genes. SEMPPR calculates the cost of completing a protein η based on the cost of the completed protein plus the expected cost of nonsense errors during translation of the said protein. For a vector of codon elongation rates $\vec{c}$:

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

Here, σ_n represents the probability that a codon sequence will translate a protein to the last codon n without a nonsense error. More generally, σ_i represents the probability that a codon sequence will translate up to and including codon i. The probability of successfully translating an individual codon i is equal to $c_i/(c_i+b)$, where b is the background nonsense error rate. Since the probability of translating up to and including codon i equals the product of the completion probabilities up to and including codon ,$\sigma_i = \prod_{k=1}^{i} c_i/(c_i + b)$. Similarly, the probability of observing a nonsense error at codon i is equal to $b/(c_i+b)$, and the probability of observing a nonsense error at codon i after translating successfully to codon i-1 is $\Pr(NSE) = \sigma_{i-1} * b/(c_i+b)$. Another term, $\xi(\vec{c})$, represents the expected cost of one nonsense error. SEMPPR calculates $\xi(\vec{c})$ by summing the cost of nonsense errors at every codon position weighted by the probability that a nonsense error will occur at that position. Parameters a_1 and a_2 represent cost of ribosome recharge and cost of peptide bond, respectively.

Nonsense errors occur when a ribosome terminates protein translation before completion, and the incomplete polypeptide is assumed non-functional, contributing an energetic cost to the cell. SEMPPR assumes a constant background nonsense error rate, so the faster a codon is translated, the lower chance it has to experience a nonsense error. Generally, a codon sequence with a higher number of fast translating codons will have a lower production cost and thus higher fitness. Since energy is invested in each peptide bond, fast-translating codons that occur near the end of a sequence have a larger impact on production cost and are subject to higher selective pressures. Using analogies to thermodynamic system state analysis (Sella and Hirsh 2005), SEMPPR calculates the probability of observing each synonymous sequence by comparing its fitness and mutation bias to that of all synonymous sequences. This can be observed as a Markov process with equilibrium:

$$f(\eta_i \mid \phi, \vec{\mu}) = \frac{\left[\prod_{k=1}^{61} (\mu_k)^{x_{k,i}}\right] e^{[-N_e q \phi \eta_i]}}{\sum_{j \in S} \left[\prod_{k=1}^{61} (\mu_k)^{x_{k,j}}\right] e^{[-N_e q \phi \eta_j]}}$$

The exponent x_k represents the number of occurrences of codon k in the given sequence, N_e represents the effective population size, a measure of genetic drift, and S represents the synonymous space, or the set of all synonymous codon sequences for a given amino acid sequence.

As previously explained, mutation values used in this experiment were taken from Gilchrist and Shah (2011) which used a maximum likelihood framework to estimate the relative mutation rates. Although the model used in the aforementioned study calculated protein cost from ribosome overhead cost rather than nonsense errors, it should still estimate accurate mutation values, because the underlying mutation process is the same for both models. This mutation model differs from simpler mutation models by the number of parameters estimated. In simpler models such as transition/transversion, there are only two estimated parameters - transition rate and transversion rate (Kimura 1980). In an even simpler model, A-T bias, there is only one estimated parameter - AT relative frequency (Hershberg and Petrov 2010). Finally in the simplest model, unbiased mutation, there are no estimated parameters. Here, there are 40 estimated parameters, one for each translating codon minus one for each synonymous codon group. For each synonymous codon group, one arbitrarily chosen mutation rate was normalized to 1, because only relative mutation rates are important in the codon-specific mutation model. Although there are 61 codons and 20 amino acids, amino acid serine was split into two synonymous groups, because the codons differed by more than one mutation, resulting in 61 codons and 21 amino acid groups.

Although the previously defined equilibrium expression returns a probability distribution of η given ϕ, Bayes' Theorem provides a posterior distribution of ϕ given η.

$$f(\phi \mid \eta_{obs}) = \frac{f(\eta_{obs} \mid \phi) f(\phi)}{\int_0^{\phi_{max}} f(\eta_{obs} \mid \phi) f(\phi) \mathrm{d}\phi}$$

Here, information known about the distribution of ϕ in the form of a prior distribution $f(\phi)$ is combined with the observed distribution of η, $f(\eta_{obs}|\phi)$. The prior distribution $f(\phi)$ is assumed noninformative where all protein production rates are equally likely. This combined information yields a posterior distribution, and its arithmetic mean is used to determine the predicted expression levels. With the codon-specific mutation model in place, SEMPPR calculated predicted production rates ϕ of 5530 verified genes from the *Saccharomyces cerevisiae* genome and compared predicted values of ϕ to empirical estimates from Bayer et al. (2004).

Results

Generated values of ϕ are found in Figure 2. Since error in predicted values is assumed lognormal, a plot of the log of predicted production values versus the log of empirical values is more informative. Results using the codon-specific mutation model correlated more strongly than results using unbiased mutation with coefficient of determination R^2 values of 0.441 and 0.426, respectively.

For low-expression genes, the results found when implementing the codon-specific mutation model indicated a slight decrease of predicted production rates while remaining relatively static in high-expression genes. Separate plots of high- and low-expression genes showed a greater improvement in the prediction of protein production rate for low-expression genes with the new mutation model. Using the codon specific mutation model, the coefficient of determination R^2 improved from .201 to .221 in the 5000 genes with the

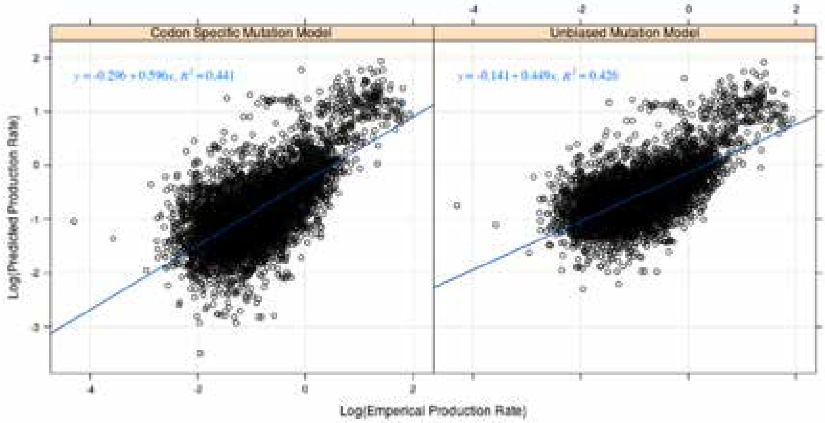

Figure 2. The first grid shows results from SEMPPR using the codon specific mutation model, and the second grid shows results from SEMPPR assuming unbiased mutation. Y-axis shows values of predicted production rate on a logarithmic scale. X-axis shows values of empirically determined production rates on a logarithmic scale. As you can see, results from SEMPPR using the codon specific mutation model are more accurate than those using unbiased mutation (Coefficient of determination0.441 vs. 0.426).

lowest expression levels, while R^2 only improved from .485 to .502 in the 530 genes with the highest expression levels. This difference indicates that the codon specific model had little effect on predicted production rates for highly expressed genes. Another interesting result lies in the difference in R^2 values between high-expression genes and low-expression genes. The higher R^2 values in high-expression genes show that SEMPPR predicts the production rate of highly-expressed genes more accurately than low-expression genes.

The indication that the codon-based mutation model affected the predicted production rate of low-expression genes more than high-expression genes is further evinced by similar deviating patterns in both graphs in genes with high expression levels. Examining the outer edges of the distribution, it is easy to identify patterns of individual data points that occur in both graphs, and these deviating patterns experienced less change in the region of highly expressed genes when implementing the new mutation model. To quantify this claim, the data were analyzed to find the percent change of ϕ for 530 genes with the highest expression rates compared to the percent change of ϕ for all other genes. On average, data points in high-expression genes changed 11.8%, while data points in low-expression genes changed 47.5%. According to these results, not only did the new mutation model conserve the overall behavior of predicted production rates in high-expression regions, but it also conserved their behavior on an individual basis.

Discussion

The increased accuracy of predicted protein production rates using the codon-specific mutation model is most likely due to more accurate prediction of the production rate of low-expression genes. Before implementing the codon-specific mutation model, SEMPPR most likely mistook codon bias seen in low-expression genes for bias due to selective pressures, when the biases were actually caused by mutation. This discrepancy occurs because several codons with high relative mutation rates also have relatively fast translation rates.

For eight of the twenty amino acids, the codon with the highest translation rate is also the codon with the highest relative mutation rate. The new codon-specific mutation model helped alleviate some error due to mutation bias. However, significant error is still present in low-expression genes according to the coefficient of determination R^2 values of the plots of low-expression genes.

Among the shortcomings of the codon-specific mutation model is the inability to distinguish the particular mechanism responsible for the heuristically determined mutation rates. In the future, researchers can develop mechanistic models that predict mutation rates, implement them in SEMPPR, and compare their results to these.

These findings emphasize the importance of mutation in codon usage bias. In low-expression genes, fitness is relatively insensitive to codon sequence, so mutation dominates the equilibrium distribution. In high-expression genes, fitness varies significantly with codon sequence, and mutation has little effect. In order to distinguish and elucidate the mechanisms behind the effects of mutation and selective pressures on codon bias, it is important that the scientific community understand both phenomena. This codon-specific mutation model improves the accuracy of predicted production rates from SEMPPR and brings the scientific community one step closer to understanding codon usage bias.

References

Akashi, H., 1994. Synonymous codon usage in Drosophila melanogaster: Natural selection and translational accuracy. Genetics 136, 927-935.

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.

Bernardi, G., Bernardi, G., 1986. Compositional constraints and genome evolution. J. Mol. Evol. 24, 111.

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

Beyer A, Hollunder J, Nasheuer HP, Wilhelm T. 2004. Posttranscriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale. Mol Cell Proteomics.

3:1083–1092.

Bulmer, M., 1988b. Codon usage and intragenic position. J. Theor. Biol. 133, 67-71.

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

Chamary, J. V., and L. D. Hurst, 2005a. Biased codon usage near intron-exon junctions: Selection on splicing enhancers, splice site recognition or something else? Trends Genet. 21: 256-259.

Chamary, J. V., and L. D. Hurst, 2005b. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol. 6: R75.1R75.12.

Chen, S. L., W. Lee, A. K. Hottes, L. Shapiro and H. H. McAdams, 2004. Codon usage between genomes is constrained by genome-wide mutational processes. Proc.Natl.Acad.Sci. USA101:3480-3485.

Comeron, J. M., 2006. Weak selection and recent mutational changes infuence polymorphic synonymous mutations in humans. Proc. Natl. Acad. Sci. USA 103: 6940-6945.

Drummond, D. A., and C. O. Wilke, 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341-352.

Gilchrist, M.A. 2007. Combining Models of Protein Translation and Population Genetics to Predict Protein Production Rates from Codon Usage Patterns. Molecular Biology & Evolution 24: 2362-2373.

Hershberg R, Petrov D. A., 2010. Evidence That Mutation Is Universally Biased towards AT in Bacteria. PLoS Genet 6(9): e1001115. doi:10.1371/journal.pgen.1001115

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.

Kimura, Motoo. 1980. A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide Sequences. Journal of Molecular Evolution 16.2 : 111-20.

Kliman, R.M., Hey, J., 1993. Reduced natural selection associated with low recombination in Drosophila melanogaster. Mol. Biol. Evol. 10,1239-1258.

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

Lin, Y. S., J. K. Byrnes, J.K. Hwang and W. H. Li, 2006. Codon-usage bias versus gene conversion in the evolution of yeast duplicate genes. Proc. Natl.Acad. Sci. USA 103: 14412-14416.

Musto, H., Romero, H., Zavala, A., 2003. Translational selection is operative for synonymous codon usage in Clostridium perfringens and Clostridium acetobutylicum. Microbiol.-SGM 149, 855-863.

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

Shah, P. and M.A. Gilchrist. 2011. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. Proceedings of the National Academy of Sciences U.S.A. 108: 10231-10236.

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.

Shields, D.C., Sharp, P.M., Higgins, D.G.,Wright, F., 1988. Silent sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. 5, 704-716.

Warnecke, T., and L. D. Hurst, 2007. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in Drosophila melanogaster. Mol. Biol. Evol. 24: 2755-2762.

Warnecke, T., N. N. Batada and L. D. Hurst, 2008 The impact of the nucleosome code on protein-coding sequence evolution in yeast. PLoS Genet. 4: 112.

About the Author

Preston Hewgley, a member of the Chancellor's Honors Program, is currently in his third year of his Pre-medical Admissions Biomedical Engineering major at the University of Tennessee. A lifelong Tennessean, he grew up near Knoxville and received his high-school diploma from Farragut High School. He currently works in the lab of Dr. Michael A. Gilchrist, a faculty member in the Department of Ecology and Evolutionary Biology at this university, researching methods in evolutionary bioinformatics to explain genomic phenomena.

About the Advisor

Michael A. Gilchrist is an Associate Professor in the Department of Ecology and Evolutionary Biology at The University of Tennessee. He is also a senior staff member at the National Institute for Mathematical and Biological Synthesis. His research interests include employing mathematical models in different biological phenomena and improving methods by which we analyze data. He is excited to be a researcher in this time that scientists are beginning to bridge the gap between models and datasets in molecular and evolutionary biology.