Model of Protein Translation with Codon Bias and Nonsense Errors

Apratim Sahay
MIT Department of Physics
(Dated: May 15, 2009)

This paper will summarize a probabilistic model(Gilchrist and Wagner [2]) of protein translation at the mRNA level that incorporates ribosome recycling and nonsense errors in a bid to shed light on codon bias. It treats translation as a probabilistic wave of ribosome occupancy travelling down the mRNA transcript. It is known that individual codons differ in both frequency of usage(codon bias) and in translation rates, consequently when codon bias is coupled with nonsense errors and ribosome recycling in their model, we observe a large effect on overall translation rate of an mRNA transcript. I will present their derivation of a simple cost function for nonsense errors, use it to derive a translational completion probability and apply it to the yeast genome. To further establish the model, they detect position dependent selection on codon bias which correlates with gene expression levels as has been empirically observed.

PACS numbers:

I. INTRODUCTION

Translation, or protein synthesis, is a process that is central to biology. The overall translation rate of a protein is known to depend on many factors: ribosome kinetics, tRNA concentrations, availability of amino acids, elongation rates of individual codons. Models of protein translation rarely take ribosome occupancy and nonsense errors into account

Research has demonstrated that synonymous codons are not translated at the same rate. Codon bias is the nonuniform usage of a particular synonymous codons (group of codons that have different nucleotide triplets but which encode the incorporation of the same amino acid during translation) within a genetic sequence. The nature and strength of codon bias varies between species as well as within a genome. Correlating codon bias to protein or mRNA abundance has been the subject of many studies. Although it is well accepted that codon bias is positively correlated with gene expression level, the importance of selection on increased translational speed or accuracy is still unclear.

One hypothesis is that codon bias results from selection to minimize the probability and associated cost 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 [4]). 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. This, in turn, leads to the prediction of increasing codon bias with codon position. The prediction of increasing codon bias with position is unique to the nonsense errors acting as a hypothesized source of selection on codon usage.

When ribosomes that have just completed translating an mRNA bind to the 5' untranslated region (UTR) of the same mRNA, then ribosomes are said to be recycled. Ribosome recycling is known to occur in eukaryotes and is thought to greatly increase the overall translational efficiency of a ribosome by reducing the mean time between the completion of translation and the next initiation event.

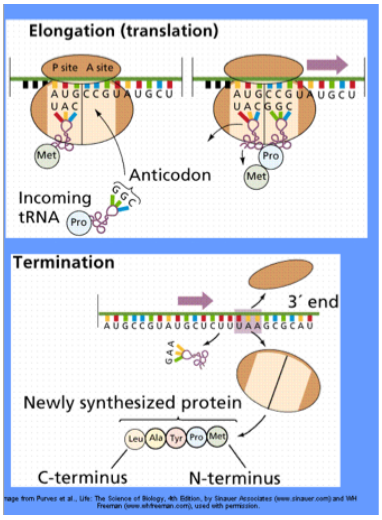

FIG. 1: From Molecular biology of cell: Translation has 3 major phases-initiation, elongation and termination. Termination may be due to a stop codon or due to a nonsense error.

Gilchrist et al present a dynamic model of protein translation that includes the phenomena of ribosome recycling and nonsense errors. This model comes in a discrete ordinary differential equation form. The model is useful in understanding the translation at the scale of an individual codon and the steady-state of the mRNA. They also explicate a continuous PDE model in their paper that is useful for understanding the large scale behavior of the system as well as the effects of ribosome recycling and nonsense errors on this behavior. I will only discuss the discrete model below.

The model allows us to calculate the probability that a nonsense error will occur for any given codon, or the probability that a ribosome will successfully complete translation of an mRNA. They are also able to derive how

translational completion probability of an mRNA transcript interacts with ribosome recycling to determine the overall translation rate of an mRNA.

II. MODEL FORMULATION

Protein translation occurs in three phases: initiation, elongation and termination. Initiation begins when a charged ribosome binds to an mRNA and ends when it translates the initial start codon. Elongation involves the interception of the correct charged tRNA by the ribosome and the transfer of the amino acid to the growing peptide chain. The third phase, termination occurs when protein elongation stops, either because a stop codon has been reached or a nonsense error has occurred. We next introduce some assumptions and parameters associated with each phase.

Initiation is known to be the rate limiting step in translation. Assume that initiation occurs with a rate γ. Ribosome recycling is incorporated by assuming that for each gene there is a fixed probability, λ, that ribosmes which complete translation are recycled back onto the same mRNA. Because it is a probability, γ can take any value between 0 and 1.

At each elongation step, the ribosome waits for a charged tRNA whose anti-codon complements the codon at the ribosome's A site. Experimental data suggest that the waiting time for the correct tRNA is the rate limiting step of the elongation process. Accordingly, the rate of elongation can vary greatly among different codons. The waiting period depends on the tRNA concentration, and we can assume that the rate of codon translations are proportional to the abundance of their cognate tRNA in the cell. Let $\vec{c}$ denote a vector of codon translation rates used during the elongation process where $\vec{c} = \{c_1, c_2, \ldots, c_n\}$, c_i is the translation rate for codon i, and n is the number of codons in the mRNA transcript.

Protein translation can terminate normally when the the translating ribosome encounters one of the 3 stop codons. It is assumed that such normal termination occurs quickly and thus this step is not explicitly modeled. Premature termination results from nonsense errors. Let us assume that the b is the nonsense error rate per unit time and is the same at all codon positions. However because the rate at which each codon is translated varies, we see the the probability that a nonsense error occurs also varies with each codon.

A. Discrete Model Formalization

Begin by defining $z_i(t)$ as the probability that a ribosome is found at codon i at time t. The vector $\vec{z}(t) = \{z_1, z_2, \ldots, z_n\}$ represents the set of probability values for all n codons in mRNA transcript. Let $t = 0$ be the time when the mRNA first becomes available for translation. Codons can become occupied by a ribosome translating the previous codon (i-1) or in the case of the first codon, by initiation. Codons can become unoccupied when the ribosome leaves the codon either by translating codon i (at rate c_i) or by disassociating with the mRNA via the termination step(with nonsense error rate b)

$$\frac{dz_i}{dt} = \begin{cases} \kappa - (\mathrm{c}_i + b)z_i, & i = 1, \\ c_{i-1}z_{i-1} - (c_i + b)z_i & i > 1. \end{cases} \tag{1}$$

with initial conditions $z_i(0)$=0 for all i. The term $\kappa(t)$ represents the total initiation rate of protein translation. Note that the negative c_i term in Eq. (1) indicates that a quickly translated codon will reduce the probability function more than a slowly translated codon.

Because we assume that the translation of the stop codon is quick, the rate of protein production at time t, $\tau(t)$, is equal to the rate at which the nth codon is translated, c_n, weighted by the probability that a ribosome is found there, $z_n(t)$, i.e

$$\tau(t) = c_n z_n(t) \tag{2}$$

Furthermore, if $\tau(t)$ is the rate of protein production and λ is the probability that a ribosome that completes translation will be recycled to the same mRNA then the rate at which ribosomes are recycled is $\tau(t)$. Thus, the initiation rate, is the sum of two processes-normal binding of free ribosomes to mRNA at rate γ and recycling:

$$\kappa(t) = \gamma + \lambda\tau(t) \tag{3}$$

If ribosome recycling occurs then initiation rate will change over time time because it is a function of the translation rate, $\tau(t)$, which is time dependent.

B. Translational completion probability

An important term that we now define is the translational completion probability of a transcript. This is the probability that a ribosome that begins translating a transcript will reach the stop codon before a nonsense error occurs. Let $\sigma(i)$ be the probability that a ribosome will complete translation up to and including codon i. From (1) it can be shown that

$$\sigma(i) = \prod_{j=1}^{i} \frac{c_j}{c_j + b} \tag{4}$$

$$\sigma(n) \approx 1 - \frac{bn}{\bar{c}}\left(1 + \frac{var(c)}{\bar{c}^2}\right) \tag{5}$$

where each term in the product $c_j/c_j + b)$ is the probabilty that a ribosome at codon j will translate the codon rather than a nonsense error occuring. $\sigma(i)$ is dependent on values in $\vec{c}$ and codon position i but is independent of time. From the definition, the probability that the entire transcript will be translated is simply $\sigma(n)$. We

simplify $\sigma(n)$ through taylor approximations about the *mean* codon translation rate of an mRNA transcript $\bar{c}$, and a further taylor series around b=0 yielding Eq. (5). It clearly shows that the the translational completion probability $\sigma(n)$, decreases with n,b and with variation in c. In contrast, $\sigma(n)$ increases with the mean codon translation rate, $\bar{c}$. This result suggests that selection for increasing $\sigma(n)$ would minimize the nonsense error rate and the variance around the mean codon translation rate while simultaneously maximizing $\bar{c}$. This minimization and maximization can only be achieved by using the fastest translating codon in each site.

C. Energetic cost of nonsense errors

If the probability of nonsense errors differs between codons, then the cost of these errors should increase with codon position because at each step of elongation, more energy is invested into building the peptide. Let us define the expected energetic cost of nonsense errors for each translational initiation event ψ for a given transcript $\vec{c}$ as

$$\psi(c) = \sum_{i=1}^{n} Pr(\text{Error at codon i})(a_1 i) = \sum_{i=1}^{n} \frac{b}{c_i}\sigma(i)(a_1 i) \quad (6)$$

where a_1 represent the energetic cost of forming a peptide bond. This cost is approximately 4 times the energy of a phosphate bond.

Suppose we switch codon c_k and c_m where $k < m$. Use $\Delta\psi_{k,m}$ to represent change in ψ caused by such a switch. Then

$$\Delta\psi_{k,m} = \left(1 - \frac{c_m}{c_m + b}\frac{c_k + b}{c_k}\right)\sum_{i=k}^{m-1}\sigma(k) \quad (7)$$

Thus we see that this switch increases or decreases the expected energetic cost of nonsense errors depending on whether $c_m > c_k$ or $c_m < c_k$ respectively. Thus, the value of ψ increases if a faster codon late in the transcript is switched with a slower translating codon earlier in the transcript. This is so because the probability of a nonsense error occurring before the faster codon is translated is lower than the slower codon and that later errors are energetically more expensive due to the larger number of peptide bonds formed.

We can also see that the the impact of a switch increases with the distance between codons switched due to the sum of σ values in Eq.(7). Furthermore, because $\sigma(i)$ is a monotonically decreasing function, switching two codons some distance apart at the start of a transcript would affect ψ more than switching two codons the same distance near the end of a transcript. This suggests that although the strength of selection on codon usage increases with codon position, the gradient of this selective force decreases with position.

III. APPLICATION TO YEAST GENOME

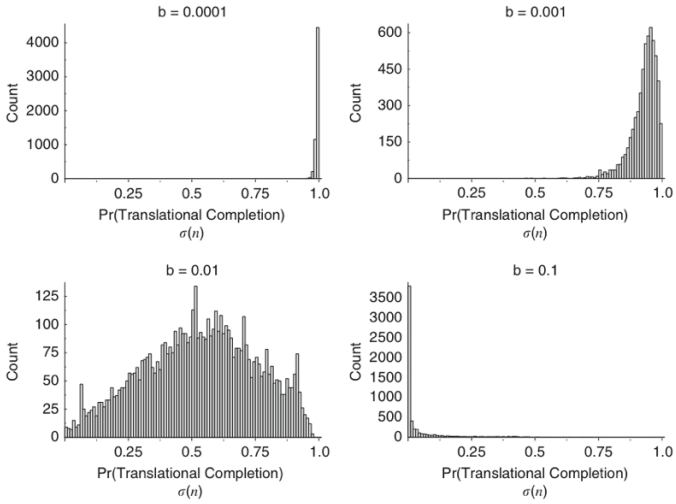

FIG. 2: From Gilchrist et al: Distribution of translational completion probabilities, $\sigma(n)$ for 5855 genes in the yeast genome for 4 different nonsense error rates, b. When b is low then all proteins have high translational probabilities., and when it is high then very few proteins are translated completely.However, in the intermediate range of error rates, the completion probabilities vary greatly among genes. Hence for a given value of b the variation in $\sigma(n)$ betweens genes is due to variation in codon usage and transcript length.

A. Calculate $\sigma(n)$ for yeast genome

To calculate the translational completion probability of an mRNA, $\sigma(n)$, we need to know (a)translation rate for each codon (b)set of codons used in mRNA (c)nonsense error rate b. We assume different values for b and calculate $\sigma(n)$ for yeast genes to show the how nonsense errors affect translational completion probabilities.

The rate limiting step for translation of an individual codon, c_i, is the rate at which ribosome intercept the correct tRNA, which is assumed to be proportional to the tRNA concentration(this is known to vary over an order of magnitude). Gilchrist et al. have a table with codon translation rates for each tRNA species.

Fig.1 shows the distribution of translational completion probabilites for 5855 genes at four different rates of nonsense errors. The mean translational completion probability ,$\sigma(n)$ at b=0.0001,0.001,.01 and 0.1 /s is 0.99,0.93,0.53, 0.047, respectively. When b is low then all proteins have high translational probabilities., and when it is high then very few proteins are translated completely.However, in the intermediate range of error rates, the completion probabilities vary greatly among genes. Hence for a given value of b the variation in $\sigma(n)$ betweens genes is due to variation in codon usage and transcript length.

B. Comparing ψ to predicted behavior in yeast genome

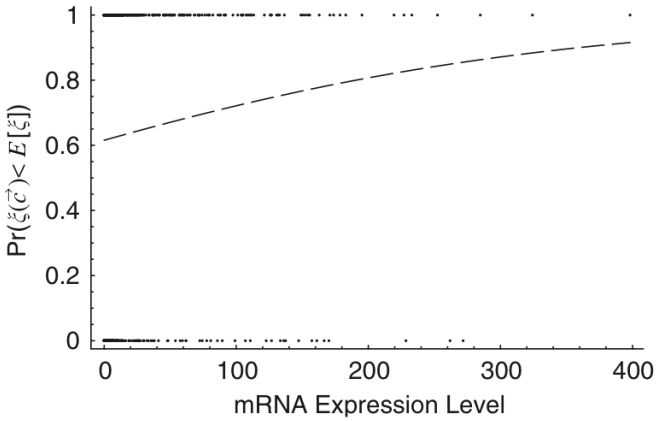

FIG. 3: From Gilchrist et al: Distribution of genes with evidence for selection on codon order vs mRNA expression level data from Beyer et all(2004). Genes were considered to have evidence for selection when their expected cost of nonsense errors of the transcript was less than the mean expected cost from a set of transcripts where codon order for each amino acid was randomized. The dashed line (–) represents the best fit line with slope=0.47, intercept=0.0048. The result indicates that strength of selection on codon usage increase with codon position and mRNA expression level.

To test the predicted properties of ψ, Gilchrist et al. take each transcript from the yeast genome and generate a null distribution of 2500 transcripts where the codons used for each amino acid were randomly rearranged. For each transcript, they calculate the expected energetic cost of nonsense errors in the null set and then compare it with the expected cost of nonsense errors, $\psi(\bar{c})$. If the observed cost was less than the mean of the null set then they score it as a 1 and if greater then 0, indicating no evidence for selection on codon order. One expects that if their model is valid and selection is acting on codon position as posited then the observed cost of nonsense errors will be less than the average cost in the null set. Indeed, of the 5855 genes examined, they find that 61% had an observed cost of nonsense errors less than the mean of the null distribution.

Another prediction was that the strength of selection on genes increase with their expression level. Consequently, we can ask whether the probability that a gene's observed cost of nonsense errors is less that expected increases with the expression level of the gene. Expression data from yeast genome studies show that the probability of finding evidence of selection on codon order increase significantly with expression level of a gene.

IV. CONCLUSIONS

The model summarized in this paper focuses on the role nonsense errors may play in shaping codon usage bias. The basic reasoning is as follows: If nonsense errors occur at a constant rate then they are more likely to occur at slow codons than at fast codons. Therefore, codon usage bias affects the probability that a nonsense error will occur along an mRNA. Gilchrist et al. model an expected nonsense error cost function to better understand the energetic cost of these errors. In this function, the assembly cost of these nonsense errors increases with the position of the codon. They indirectly test this cost function and the translation model on which it is based by asking whether they can detect a response to an increasing selection gradient on codon usage bias.

Ribosome recycling is likely to have evolved as a means of increasing the translational efficiency of the ribosome population. However ribosome recycling can only occur if the ribosome translates the entire transcript. Thus the overall efficacy of ribosome recycling is limited by the translational probability of a transcript, which in turn is determined by the probability of a nonsense error at one of the codons.

By tying together translational efficiency with a very insightful discussion of translational completion probability and nonsense errors, Gilchrist et al leave us with a clearer understanding of how codon usage bias underlies this interaction.

Acknowledgments

I wish to acknowledge my undergraduate thesis advisor Philippe Cluzel for introducing me to the rich topic of codon bias. I thank Professors Kardar and Mirny for a thoroughly stimulating class.

[1] Eyre-Walker A Akashi H. Translational selection and molecular evolution. *Curr. Opin. Genetic. Dev, 11, 660-666*, 2003.

[2] Wagner A. Gilchrist MA. A model of protein translation including codon bias, nonsense errors, and ribosome recycling. *Journal of Theoretical Biology 239, 412-434*, 2006.

[3] Bulmer M. Codon usage and intragenic position. *Journal of Theoretical Biology 133, 67-71*, 1988.

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

[5] Chou T. Ribosome recycling, diffusion, and mrna loop formation in translational regulation. *Biophysical Journal, 85, 755-773*, 2003.

[6] Ikemura T. Correlation between abundance of e. coli transfer rna's and codon frequency in genome. *J.Mol Bio, 151, 389-409*, 1981.

[7] Zubay G Zhang SP, Goldman E. Clustering of low usage codons and ribosome movement. *Journal of theoretical biology, 170, 339-354*, 1994.

[1–7]