Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Theoretical Biology 239 (2006) 417–434

Journal of Theoretical Biology

www.elsevier.com/locate/yjtbi

A model of protein translation including codon bias, nonsense errors, and ribosome recycling

Michael A. Gilchrist[a,*], Andreas Wagner[b]

[a]*Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN 37996, USA*
[b]*Department of Biology, University of New Mexico, Albuquerque, NM 87131, USA*

Received 18 January 2005; received in revised form 5 August 2005; accepted 8 August 2005
Available online 19 September 2005

Abstract

We present and analyse a model of protein translation at the scale of an individual messenger RNA (mRNA) transcript. The model we develop is unique in that it incorporates the phenomena of ribosome recycling and nonsense errors. The model conceptualizes translation as a probabilistic wave of ribosome occupancy traveling down a heterogeneous medium, the mRNA transcript. Our results show that the heterogeneity of the codon translation rates along the mRNA results in short-scale spikes and dips in the wave. Nonsense errors attenuate this wave on a longer scale while ribosome recycling reinforces it. We find that the combination of nonsense errors and codon usage bias can have a large effect on the probability that a ribosome will completely translate a transcript. We also elucidate how these forces interact with ribosome recycling to determine the overall translation rate of an mRNA transcript. We derive a simple cost function for nonsense errors using our model and apply this function to the yeast (*Saccharomyces cervisiae*) genome. Using this function we are able to detect position dependent selection on codon bias which correlates with gene expression levels as predicted a priori. These results indirectly validate our underlying model assumptions and confirm that nonsense errors can play an important role in shaping codon usage bias.

Keywords: Protein translation; Codon bias; Nonsense errors; Ribosome recycling; Translational efficiency; Translational accuracy

1. Introduction

In this study we explicitly model how codon usage bias and nonsense errors affect the probability that a ribosome will successfully translate an messenger RNA (mRNA) transcript. We also determine how this probability interacts with the process of ribosome recycling to determine the overall translation rate of a protein. Our model is one in a long tradition of models of protein translation (Bergmann and Lodish, 1979; Harley et al., 1981; Menninger, 1983; Liljenström and von Heijne, 1987; Bulmer, 1991; Zhang et al., 1994; Chou, 2003), but is the first to consider both nonsense errors and ribosome recycling.

Codon bias is the non-random usage of synonymous codons within a gene (Ikemura, 1981; Bennetzen and Hall, 1982; Sharp and Li, 1987). It has been extensively documented across a wide range of organisms and varies greatly both between DNA sequences within a genome and between species (e.g. Ikemura, 1981, 1982, 1985; Bennetzen and Hall, 1982; Sharp and Li, 1987; Ghosh et al., 2000; Carbone et al., 2003; Mougel et al., 2004). Most explanations of codon bias generally involve a mixture of factors including purely physical forces, such as mutational bias and recombination, and selection for increased translational efficiency or accuracy (e.g. Bernardi and Bernardi, 1986; Bulmer1988a, 1991; Shields et al., 1988; Kliman and Hey, 1993, 1994; Akashi, 1994, 2003; Xia, 1996; Akashi and Eyre-Walker, 1998; Xia, 1998; Musto et al., 1999; McVean and Charlesworth, 1999; Ghosh et al., 2000; Wagner, 2000;

[*]Corresponding author. Tel.: +1 865 974 6453; fax: +1 865 974 3065.
E-mail address: mikeg@utk.edu (M.A. Gilchrist).

0022-5193/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2005.08.007

Birdsell, 2002; Comeron and Kreitman, 2002; Musto et al., 2003). In general, these explanations ignore the role of nonsense errors. However, there are a few notable exceptions (Menninger, 1983; Eyre-Walker, 1996; Berg and Silva, 1997; Hooper and Berg, 2000; Qin et al., 2004).

Ribosome recycling occurs when a ribosome which has just completed translating an mRNA binds to the 5′ untranslated region (UTR) of the same mRNA. Such recycling is made possible in eukaryotes by the loop-like arrangement of an mRNA and its translation initiation complex (Jacobson, 1996; Sachs, 2000; Welch et al., 2000; Kapp and Lorsch, 2004). In contrast, prokaryotes lack such an arrangement and, consequently, ribosome recycling is thought to be uncommon in prokaryotes. For individual eukaryotic genes, experimental and theoretical evidence indicates that ribosome recycling contributes significantly to its protein production rate (Gallie, 1991; Niepel et al., 1999; Khaleghpour et al., 2001; Chou, 2003; Rajkowitsch et al., 2004). In terms of the ribosome population within a cell, recycling should greatly increase the overall translational efficiency of each ribosome in the population by reducing the mean time between the completion of translation and the next initiation event.

While recycling can increase the overall translational efficiency of a ribosome which completely translates an mRNA, a significant proportion of ribosomes terminate translation before they reach the final stop codon (Manley, 1978; Tsung et al., 1989; Jorgensen and Kurland, 1990; Kurland, 1992). These premature termination events are called nonsense errors (also referred to as processivity errors) and include ribosome drop-off, improper translation of release factors, and frameshifts (Kurland, 1992; Hooper and Berg, 2000). Many of the incomplete peptides resulting nonsense errors will be non-functional, possibly toxic to the cell (Menninger, 1978) or, alternatively, can tie up essential cell resources such as tRNAs (Dincbas et al., 1999; Cruz-Vera et al., 2004). In order to avoid these problems, incomplete peptides need to be recognized and broken down by the cell. The production and breakdown of these incomplete peptides may represent a significant energetic cost to the cell, especially for highly expressed proteins. Consequently, nonsense errors are likely to be a potent source of selection shaping the evolution of the protein translational process (Kurland, 1992; Eyre-Walker, 1996; Hooper and Berg, 2000).

In this study we present and analyse a dynamic model of protein translation that includes the phenomena of ribosome recycling and nonsense errors. This model comes in two versions, a discrete ordinary differential equation version and a continuous partial differential equation version. The discrete model is particularly useful for understanding translation at the scale of an individual codon and the steady-state of the mRNA. The continuous model is useful for understanding the larger scale behavior of the system as well as the effects of ribosome recycling and nonsense errors on this behavior.

Our model allows us to calculate the probability that a nonsense error will occur for any given codon. In addition, our model illustrates how the risks of nonsense errors compound one another to determine the probability that a ribosome will successfully complete translation of an mRNA. From this work we are also able to derive how the translational completion probability of an mRNA transcript interacts with ribosome recycling to determine the overall translation rate of an mRNA.

To indirectly test our model's underlying assumption, we use our discrete version to build a simple function that describes the expected cost of nonsense errors for a given transcript. The cost function we derive makes explicit how the cost of a nonsense error should increase with codon position. We test our model by asking whether the expected cost of nonsense errors, when compared to a null set of transcripts built with the same set of codons, conforms to the patterns predicted by Eyre-Walker (1996).

In addition to its applications for understanding codon usage patterns, our model has many potential uses, some of which we develop elsewhere. For example, we are currently working on a manuscript (Gilchrist and Wagner, in preparation) which shows how to infer protein translation rates from measurements of ribosome densities per mRNA published by Arava et al. (2003). This application illustrates how mechanistic models of biological processes like the one we develop here can serve to integrate a wide variety of genome-scale biological information (Troyanskaya et al., 2003; Jansen et al., 2003; Gilchrist et al., 2004; Lee et al., 2004; Beyer et al., 2004).

2. Model motivation and formulation

In this section we state and motivate our model assumptions. We then formalize these assumptions into a set of coupled differential equations which comprise a discrete, time dependent model of protein translation.

Protein translation occurs in three distinct phases: initiation, elongation, and termination. The first phase, initiation, begins when a charged ribosome binds to an mRNA and ends when it translates the initial start codon. The second phase, peptide elongation, occurs after initiation and 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 and final phase, translation termination, occurs when protein elongation stops, either because a stop codon has been reached or a nonsense error has occurred.

2.1. Initiation

The initiation of protein translation occurs when a ribosome binds to the 5′ UTR of an mRNA and then moves along the mRNA until it intercepts the appropriate start codon. Theoretical and empirical studies suggest that initiation is the 'rate limiting' step in the translation process (Bergmann and Lodish, 1979; Liljenström and von Heijne, 1987; Varenne et al., 1984).

Initiation can occur de novo, but in eukaryotes it can also occur through ribosome recycling. In de novo initiation, the initiating ribosome comes from the pool of free ribosomes floating in the cytosol. Here we assume that in any given physiological condition and for any given gene there is a rate, γ, at which such de novo initiation occurs. We incorporate ribosome recycling, into our model by assuming that for each gene there is a fixed probability, λ, that ribosomes which complete translation are recycled back onto the same mRNA. Because it is a probability, λ can take any value between 0 and 1. Ribosome recycling is thought to be common in eukaryotes due to the circular nature of the transcript when bound to the translation complex. Experimental evidence suggests that the probability can be at least 50% and is affected by the stability of the secondary structure of the 5′ UTR of the mRNA (Niepel et al., 1999; Rajkowitsch et al., 2004). Quantitative information on γ is poor and estimates of λ are essentially unknown. Nonetheless, given the variation in 5′ and 3′ UTRs, both parameters are likely to vary from one gene to another (Kozak, 2002).

2.2. Elongation

Once initiation has been completed the ribosome begins the elongation stage of translation. Because initiation is assumed to be the rate limiting step, and because recent transcriptome-scale empirical data indicate that ribosome densities per codon are generally quite low (Arava et al., 2003), we ignore any ribosome–ribosome interference during the elongation process. At each elongation step, the ribosome waits until it intercepts a charged tRNA whose anti-codon pattern complements the codon at the ribosome's A site. Once the ribosome intercepts the correct tRNA, the amino acid is transferred from the tRNA to the peptide chain associated with the ribosome, and the ribosome ratchets forward one codon. Experimental data suggests that the waiting time for the correct tRNA is the rate limiting step of the elongation process (Bergmann and Lodish, 1979; Liljenström and von Heijne, 1987; Varenne et al., 1984). This is in accord with observations that the rate of elongation can vary greatly among different codons (Gouy and Grantham, 1980; Chavancy and Garel, 1981; Varenne et al., 1984; Thomas et al., 1988; Curran and Yarus, 1989; Sorensen et al., 1989). The waiting period, in turn, depends on the tRNA's concentration. For non-wobble codons, we assume that the rates of codon translations are proportional to the abundance of their cognate tRNA within a cell. In vitro and in vivo studies with *Escherichia coli* indicate that wobble tRNAs translate different codons at different rates (Thomas et al., 1988; Curran and Yarus, 1989). Based on Curran and Yaruss (1989) measurements, we reduced the translation rates of G–U and I–C wobble codons by 39% and 36% relative to G–C and I–U ending codons, respectively. The specific rates used for each codon can be found in the supplemental Table S1. Rates were scaled such that the average translation rate of all of the codons was 10 amino acids/s.

More complicated models of the elongation process that include such things as initial binding, codon recognition, GTP hydrolysis, and kinetic proofreading, could be used to produce more refined estimates of translation rates of different codons. Ideally, models of these intra-ribosomal processes would be nested within our current framework. However, such extensions are outside of the scope of our current study.

We will use $\vec{c}$ to 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. Note that $\vec{c}$ does not include the start and stop codons, which we consider to be part of the initiation and termination processes, respectively.

2.3. Termination

Protein translation can either terminate normally with the ribosome completing the translation of the mRNA transcript or prematurely when a nonsense error occurs. In normal translation termination, a polypeptide is released after the translating ribosome encounters one of three possible stop codon and its corresponding release factor. Note that a nonsense error which leads to termination at a stop codon is functionally not an error since the released product is a complete peptide. For simplicity, we assume that termination occurs quickly and, as a result, we do not explicitly model this step in the translation process. Because the stop codon is only one out of hundreds or thousands of translated codons, modeling this step explicitly would not affect our results in any noticeable manner.

Premature translation termination results from nonsense errors. Such errors can have multiple causes, such as reading frameshifts, ribosome drop-off, or false termination (Kurland, 1992; Hooper and Berg, 2000). As has been explicitly and implicitly assumed in other studies (e.g. Curran and Yarus, 1989; Thomas et al., 1988), we assume that the nonsense error rate per unit time is the same at all codon positions. In spite of this assumption, we will show that the probability that a nonsense error occurs varies with each type of codon. This is because the rate at which each codon type is translated varies. In addition, because nonsense errors often result in the ribosome disassociating from the mRNA, we also assume that nonsense errors prevent the possibility of a ribosome being recycled on the same mRNA.

2.4. Formalization

We will now formalize the assumptions outlined above into a discrete mathematical model of protein translation along a focal mRNA. All model parameter definitions can be found in Table 1. We begin by defining $z_i(t)$ as the probability that a ribosome is found at codon i at time t of the focal mRNA. The term $\vec{z}(t) = \{z_1(t), z_2(t), \ldots, z_n(t)\}$ represents the set of probability values for all n codons involved in elongation within the mRNA transcript.

Because we are focusing on the translation processes at the level of a single mRNA transcript, we define $t = 0$ as the time at which an 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, where $i = 1$, by initiation. Codons become unoccupied when the ribosome leaves the codon either by translating codon i or by disassociating with the mRNA via a ribosome drop-off or other nonsense error. Reminding the reader that c_i represents the translation rate of the ith codon and b represents the nonsense error rate, we write:

$$\frac{\mathrm{d}z_i}{\mathrm{d}t} = \begin{cases} \kappa - (c_i + b)z_i, & i = 1, \\ c_{i-1}z_{i-1} - (c_i + b)z_i, & i > 1 \end{cases} \tag{1}$$

with the initial conditions

$$z_i(0) = 0 \quad \text{for all}\, i. \tag{2}$$

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.

As previously mentioned, experimental evidence (Niepel et al., 1999; Rajkowitsch et al., 2004) suggests that in eukaryotes, ribosomes that complete translation have a significant probability of reattaching to the 5′ end of the same mRNA. Thus $\kappa(t)$ is the sum of two separate processes, the binding of free ribosomes to the mRNA (de novo initiation), γ, and the recycling of ribosome which have just completed translating the nth codon.

Because we assume that translation of stop codons 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 of a ribosome being found there, $z_n(t)$, i.e.

$$\tau(t) = c_n\, z_n(t). \tag{3}$$

Further, if $\tau(t)$ is the rate of protein translation and λ is the probability that a ribosome that completes translation will be recycled to the 5′ UTR of the same mRNA, then the rate at which ribosomes are recycled $\lambda\,\tau(t)$. Thus,

$$\kappa(t) = \gamma + \lambda\,\tau(t). \tag{4}$$

If ribosome recycling occurs, i.e. $\lambda > 0$, then the initiation rate will change over time because it is a function of the translation rate, $\tau(t)$, which is time dependent.

Table 1
List of parameters and variables in formulation of the discrete and continous models. Steady-state values of variables are indicated by a ^

c_i	Translation rate of codon i
$\vec{c}$	Set of c_i values for an entire mRNA sequence
$z_i(t)$	Ribosome occupancy probability, i.e. the probability that a ribosome is found at codon i of an mRNA sequence at time t
$\vec{z}$	Set of $z_i(t)$ values for an entire mRNA sequence
n	Number of codons in the mRNA
b	Nonsense error rate
$\sigma(i)$	Probability a ribosome that begins translating a sequence will translate up to and including codon i
$\sigma(n)$	Translational completion probability of a sequence, i.e. the probability of a ribsome will completely translate the sequence
γ	The rate at which new (i.e. non-recycled) ribosomes bind to the 5′ UTR of mRNA
λ	Probability a ribosome which completes translation is recycled back to the 3′ UTR
$\kappa(t)$	The total rate at which ribosomes bind to the 5′ UTR at time t
$\tau(t)$	Protein translation rate at time t
$\xi(\vec{c})$	Expected cost of nonsense errors for a given sequence $\vec{c}$
$\Delta\xi_{k,m}$	Change in $\xi(\vec{c})$ value when codons k and m are switched
$u(t,x)$	Density of ribosomes at location x at time t (continuous model)
s	the effective codon translation rate of a sequence (continuous model)
$r(t)$	Number of waves of translation that have been completed by time t (continuous model)

2.5. Defining the energetic cost of nonsense errors

If the probability of nonsense errors differs between codons and the peptide product of these errors is generally non-functional, then the cost of these errors should increase with codon position (Kurland, 1992; Eyre-Walker, 1996; Akashi, 2001; Hooper and Berg, 2000; Qin et al., 2004). This is because at each step of elongation, more energy is invested into the building of the peptide. We define the expected energetic cost of nonsense errors for each translational initiation event, ξ, for a given transcript $\vec{c}$ as

$$\xi(\vec{c}) = \sum_{i=1} n \Pr(\text{Nonsense error at codon } i)\,(a_1 + a_2 i) \tag{5}$$

where a_1 represent the energetic costs of recharging the ribosome and a_2 represents the energetic cost of forming a peptide bond. These costs are approximately 2 and 4 high energy phosphate bonds, P, respectively (Bulmer, 1991; Wagner, 2005). Note that Eq. (5) only includes the cost of peptide assembly and does not include other potential costs such as the toxic effects of incomplete peptides.

We can test whether evidence for selection to reduce the cost of nonsense errors exists by comparing the expected energetic cost of nonsense errors of an observed transcript in yeast to a set of hypothetical transcripts in which codon order for each amino acid has been randomized separately. This approach allows us to change the ordering of codons without changing the amino acid sequence of a gene or its overall codon bias. For any given transcript we will refer to its rearranged forms as its null set.

A priori we would expect the observed costs for a transcript to be consistently smaller than the mean costs of the transcript's null set. Because ξ is a cost per initiation event, we also expect that more highly expressed genes, generally those with greater mRNA abundances, would show greater evidence of selection for reducing the observed ξ relative to its null set. We will take the occurance of these patterns as support for the validity of our basic model of protein translation, its underlying assumptions, and our formulation of the cost function ξ.

3. Results

While the formulation of our discrete codon model is rather straightforward, the analysis of its dynamics as it approaches the steady-state is not. This is due to the discrete and heterogeneous nature of the system. Its discreteness requires that we have a system of equations which includes one equation for each of the n codons in an mRNA transcript. The heterogeneity in codon translation rates prevents any direct simplification of these coupled equations. Therefore, while dynamic solutions of our model can be calculated using standard numerical techniques, these calculations are computationally intensive.

To elucidate the behavior of this system, we take the following two steps. First, we derive a continuous approximation to the discrete codon model. This continuous approximation results in a conceptually and mathematically concise model that can be solved analytically. Our solution shows that the probability that a ribosome occupies a codon can be viewed as a traveling wave that moves along the mRNA transcript. Because of nonsense errors, the wave decays as it moves. However, because of ribosome recycling the wave can reinforce itself as well. This solution also helps us understand how the system approaches its steady-state which we expect to be realized for long-lived mRNAs that are translated many times. In a second step, we derive the analytic solution to this steady-state using the original discrete model.

3.1. Translational completion probability: $\sigma(n)$

One common factor of all of the approaches we take is their dependence on a term we now define as 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. Thus, we begin our analysis by exploring this function and its sensitivity to a number of underlying model parameters.

Let $\sigma(i)$ represent the probability that a ribosome will complete translation up to and including codon i. From (1) it can be shown that,

$$\sigma(i) = \prod_{j=1}^{i} \frac{c_j}{c_j + b}, \tag{6}$$

where each term in the product function $c_j/(c_j + b)$ represents the probability that a ribosome at codon j will translate the codon as opposed to a nonsense error occurring. If $b > 0$ then $\sigma(i)$ is a strictly decreasing function of i and is less than one for all i. Setting $b = 0$ implies that $\sigma(i) = 1$ for all values of i and results in a model conceptually similar to those which ignore nonsense errors (e.g. Liljenström and von Heijne, 1987).

It is worth noting that codon translation probability, $\sigma(i)$, is dependent on the values in $\vec{c}$ and codon position i but independent of time, t. Based on our definition of $\sigma(i)$, it follows that the probability that a ribosome will translate an entire codon transcript of length n is simply equal to $\sigma(n)$. Therefore, $\sigma(n)$ represents the *translational completion probability* of a transcript. Eq. (6) indicates that $\sigma(n)$ is affected by the nonsense error rate, b, the set of codon translation rates in the transcript, $\vec{c}$, and the total number of codons in the transcript, n. We can better understand how these factors affect $\sigma(n)$ through a set of Taylor Series approximations.

3.2. Understanding $\sigma(n)$ through approximation

In this section we show how we can better understand the relationship between the translational completion probability of a transcript, $\sigma(n)$, and its parameters using Taylor Series approximations of $\sigma(n)$. We begin by taking a second order Taylor Series approximation of $\sigma(n)$ about the mean codon translation rate of an mRNA transcript, $\bar{c}$. Doing so yields,

$$\sigma(n) = \prod_{i=1}^{n} \frac{c_i}{c_i + b} \approx \left(\frac{\bar{c}}{\bar{c}+b}\right)^n \times \exp\left[-\frac{n}{2}\text{var}(\vec{c})\left(\frac{1}{\bar{c}^2} - \frac{1}{(\bar{c}+b)^2}\right)\right], \tag{7}$$

where var($\vec{c}$) represents the variance of the translation rates in $\vec{c}$. The higher order moments of the distribution of $\vec{c}$, such as skew and kurtosis, become important in higher order approximations.

Given that the nonsense error rate, b, is likely to be much smaller than the mean codon translation rate, $\bar{c}$, we can further approximate the translational completion probability of a transcript, $\sigma(n)$, by taking a first order Taylor Series approximation of $\sigma(n)$ in (7) about $b = 0$. Doing so yields,

$$\sigma(n) \approx 1 - \frac{b\,n}{\bar{c}}\left(1 + \frac{\text{var}(c)}{\bar{c}^2}\right). \tag{8}$$

Eq. (8) clearly indicates that the translational completion probability, $\sigma(n)$ decreases with n, b, and with variation in $\vec{c}$. In contrast $\sigma(n)$ increases with the mean codon translation rate, $\bar{c}$.

The relationship between variation in $\vec{c}$ and the translational completion probability of a transcript, $\sigma(n)$, can be simplified further by noting that the harmonic mean of $\vec{c}$, $\bar{c}_H$, can be approximated to the second order around $1/c$ as $\bar{c}_H = (1/\bar{c} + \text{var}(c)/\bar{c}^3)^{-1}$. Using this second approximation we find that,

$$\sigma(n) \approx 1 - \frac{b\,n}{\bar{c}_H}. \tag{9}$$

Eq. (9) demonstrates that the effect of variation in $\vec{c}$ on σ is similar to the effect that such variation has on the harmonic mean of $\vec{c}$.

This result will re-present itself when we explore the continuous approximation of our discrete codon model and suggests that selection for increasing $\sigma(n)$ can manifest itself by maximizing the harmonic mean of $\vec{c}$, $\bar{c}_H$. If there is selection for maximizing the translational completion probability of an mRNA transcript, $\sigma(n)$, then we would expect selection to minimize the nonsense error rate, b, and the variance around the mean codon translation rate, var($\bar{c}$), while, simultaneously, maximizing $\bar{c}$. This minimization of var($\bar{c}$) and maximization of $\bar{c}$ can be achieved by using only the fastest translating codons at every site.

3.3. Calculating $\sigma(n)$ for the yeast genome

While approximating the translational completion probability of an mRNA, $\sigma(n)$, is useful for understanding the factors that shape it, if we know (a) the translation rate of each codon, (b) the set of codons used in an mRNA, and (c) the nonsense error rate, b, we can calculate $\sigma(n)$ for a gene exactly. We next calculate $\sigma(n)$ under different assumed values of b for all confirmed genes in the yeast genome. We do this to illustrate the potential importance of nonsense errors for translational completion probabilities.

The rate limiting step for the translation of an individual codon, c_i, is the rate at which ribosomes intercept the correct tRNA. This interception rate should be proportional to the tRNA's concentration in the cell. The concentration of tRNAs in a cell has been measured or can be estimated in a number of different organisms and has been shown to vary over an order of magnitude (Ikemura, 1982, 1985; Percudani et al., 1997; Akashi, 2003). We calculate a proportionality constant between relative tRNA abundance and codon translation rates such that the average codon translation rate in yeast is 10 amino acids/s. This rate was chosen based on the observation that ~85% of ribosomes (rib) are involved in protein translation during the exponential growth phase (Arava et al., 2003) and that the efficiency of protein translation per ribosome in the cell is 8.8 amino acids/s (aa/s) (Tuite, 1989) ($8.8\,\text{aa/s} \div 0.85\,\text{rib} \approx 10\,\text{aa/(s rib)}$). The resulting codon translation rates for each tRNA species are listed in Supplemental Table S1.

In the yeast genome there are currently 5889 confirmed protein coding genes. We ignored 34 of these confirmed transcripts due to the fact that they have internal stop codons. Fig. 1 shows the distribution of translational completion probabilities for the remaining 5855 genes at four different rates of nonsense errors.

The mean translational completion probability, $\sigma(n)$, of the yeast genome at $b = 0.0001$, 0.001, 0.01, and 0.1/s is 0.99, 0.93, 0.53, and 0.047, respectively. In general, when the nonsense error rate is low, e.g. $b = 0.0001$, all proteins have high protein translation probabilities. In contrast, when the nonsense error rate is high, e.g. $b = 0.1$, very few proteins are translated completely. At both of these extremes, codon usage and transcript length have little impact on $\sigma(n)$. However, in the intermediate range of nonsense error rates, protein translation probability values vary greatly among genes, as the figure shows. Our approximation of $\sigma(n)$ in (7) shows that for a given value of b we can attribute the

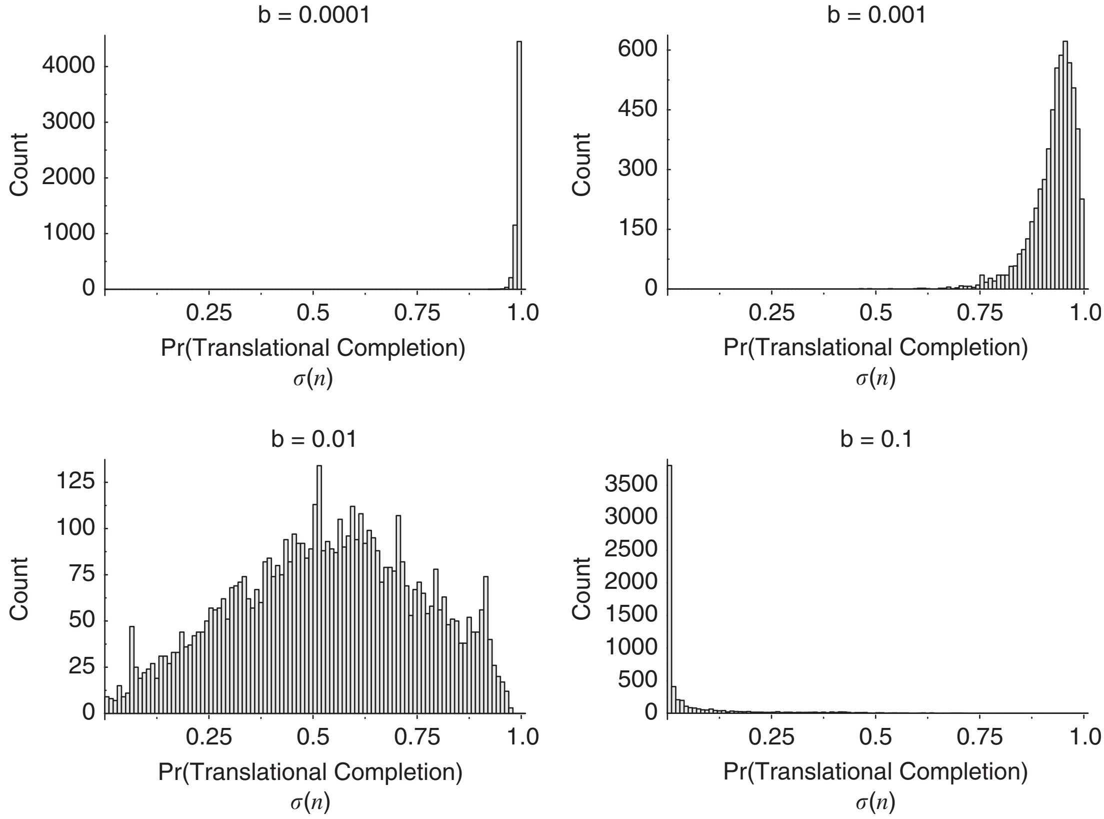

Fig. 1. Distribution of translational completion probabilities, $\sigma(n)$, for 5855 confirmed genes in the yeast genome for various nonsense error rates, b. The translation rate for a particular codon was inferred from tRNA densities within a cell and by assuming that the mean translation rate across all tRNA species was 10 amino acids/s (see text for more details).

Table 2
Table of estimate nonsense error rates and the mean protein translation probability for the yeast genome based on these values. Error rates were calculated assuming a uniform codon translation rate of 10 aa/s

Nonsense error rate b	Organism	Type of observation	Mean $\sigma(n)$ for yeast	Source
0.0025	*E. coli*	Direct	0.84	Jorgensen and Kurland (1990)
0.0052			0.70	Mean of Tsung et al. (1989) and Jorgensen and Kurland (1990)
0.0078	*E. coli*	Direct	0.60	Tsung et al. (1989)
0.0604	Yeast	Indirect	0.10	Arava et al. (2003)

variation in $\sigma(n)$ between genes to variation in codon usage and transcript length.

With this insight into how nonsense error rates, b, affect the distribution of the translational completion probabilities, we examine the experimental data that might inform us of b's actual value. Available estimates of b and their corresponding mean protein translation probability for the yeast genome are presented in Table 2. Currently, two direct estimates of the frequency of nonsense errors can be found in the literature (Tsung et al., 1989; Jorgensen and Kurland, 1990). Unfortunately, both of these estimates are based on observations in *E. coli*. Arava et al. (2003) indirectly estimate a nonsense error frequency for yeast. However, this rate is significantly greater than the values for *E. coli* and results in an average protein translation probability of only 0.10 for yeast. Clearly the estimate by Arava et al. (2003) is too large to be biologically feasible. Consequently, for the rest of our study we will use the mean of estimates from Jorgensen and Kurland (1990) and Tsung et al. (1989), $b = 0.00515$.

3.4. Dynamic behavior

While the generation of numerical solutions to our model is computationally straightforward understanding the behavior of the model from such solutions is less so. However, it is possible to derive a continuous model which is analogous to the discrete model. This approximation can be solved analytically, leading to a great deal of insight into the behavior of both the continuous and discrete forms of our model. As a result we continue our study of the discrete codon model by deriving and assessing the analytic solution to its continuous approximation. Using the insight gained from our analysis of the continuous approximation, we will then determine numerical solutions of the discrete codon model. To do so, we use the routine NDSolve that is a part of Mathematica v5.0 (Wolfram Research Inc., 2003).

3.5. Continuous approximations

We now present a partial differential equation which represents a continuous approximation to our discrete set of Eqs. (1)–(3). Because we replace an mRNA transcript with its set of discrete heterogeneous translation rates with one with a uniform translation rate and we treat codon position as a continuous, rather than a discrete, variable, this approximation allows us to generate analytic solutions.

Let $u(t,x)$ be the probability that a ribosome is centered at point x where $0 \geqslant x \geqslant n$. The spatial variable x corresponds to the codon number i in the discrete model. The variables i and x are measured in the same units: codons. The fundamental difference between i and x, however, is that x is not restricted to only integer values. (It would be possible to scale x so that it ranges between 0 and 1, representing the first and last codons, respectively. However, doing so would make comparisons of translational dynamics between genes of different lengths more difficult.)

Based on our discrete codon model, the behavior of $u(t,x)$ follows the following PDE:

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} s = -bu, \tag{10}$$

where

$$s = -\frac{nb}{\ln(\sigma(n))}. \tag{11}$$

The initial and renewal conditions for (10) are,

$$u(0,x) = 0, \tag{12}$$

and

$$u(t,0) = \gamma/s + \lambda u(t,n), \tag{13}$$

respectively.

The system of Eqs. (10)–(13) is completely analogous to the system of equations in the discrete model, (1)–(3). For example, Eq. (10) essentially states that the probability of a ribosome occurring at a particular point moves along the mRNA moves along the transcript at a constant rate s. Furthermore, as the wave of probability moves along the transcript it decays at a constant rate b over time. The initial condition reflects the fact that in the beginning the probability of finding a ribosome anywhere along the transcript is zero. The renewal condition states that the rate at which ribosomes are flowing into the system is equal to their initiation rate, scaled by the rate at which they move along the transcript, and the rate at which ribosomes that complete translation are recycled.

The term s represents the *effective* uniform codon translation rate for the entire mRNA transcript. It is analogous to the average rate at which ribosomes move along the transcript adjusted for the fact that some ribosomes terminate protein translation early. Even in the limiting case where, $b = 0$, s is not simply the arithmetic mean of $\vec{c}$. This is because ribosomes spend different amounts of time at different codons. Instead, one can show that,

$$\lim_{b\to 0} s = \frac{n}{\sum_{i=1}^{n}(1/c_i)} = \frac{1}{\bar{c}_H}. \tag{14}$$

Eq. (14) indicates that the actual average rate at which a ribosome travels along the transcript is equal to the inverse of the harmonic mean of $\vec{c}$, $\bar{c}_H$. The dependence on the harmonic mean rather than the arithmetic mean results from the fact that the wait time for each elongation step is equal to $1/c_i$ and is consistent with our earlier approximations of $\sigma(n)$ derived from the discrete model.

We can solve (10) given its renewal and initial conditions in (13) and (12) using the method of characteristics (Murray, 1993). Using this technique we first discover that,

$$u(t,x) = u(t - x/s, 0)\sigma(n)^{x/n}. \tag{15}$$

The term $u(t - x/s, 0)$ represents the probability that a ribosome would have been found at the start of the transcript x/s time units ago.

The term $\sigma(n)^{x/n}$ represents the decay in the probability that a ribosome is still associated with the mRNA (i.e. that a nonsense error has not occurred) by the time it is expected to reach point x. Eq. (15) indicates that the probability of finding a ribosome at point x at time t, $u(t,x)$, is equal to the probability of a ribosome being found at the time when this wave originated at $x = 0$, i.e. $u(t - x/s, 0)$, weighted by $\sigma(n)^{x/m}$.

From (15), we see that $u(t,x)$ is proportional to the density of ribosomes at the beginning of the transcript, i.e. $u(t,0)$. The method of characteristics also shows that

$$u(t,0) = \frac{\gamma}{s}\sum_{i=0}^{r(t)}(\lambda\sigma(n))^i \tag{16}$$

where

$$r(t) = \left\lfloor t\,\frac{s}{n}\right\rfloor. \tag{17}$$

The notation $\lfloor\,\rfloor$ represents the floor function (which is sometimes referred to as the greatest integer function).

The variable r represents the number of waves of translation which have been completed. Note that for $x > ts$ the position x is too far downstream for any ribosomes to have arrived given the rate at which ribosomes move along the transcript. Under this condition, $r(t - x/s) = -1$ and, consequently, $(\gamma/s) \sum_{i=0}^{r} (\lambda\,\sigma(n))^i = 0$, i.e. the probability of finding a ribosome at these positions is zero (Fig. 2 when $t \leqslant 100$). This behavior is consistent with our initial conditions in (12).

Eq. (16) indicates that when ribosomes are recycled, i.e. when $\lambda > 0$, the density of ribosomes at the beginning of the transcript, $u(t, 0)$, increases over time. Furthermore, this term increases in a step-like manner with each completed wave of translation. The size of each successive wave is determined by r, $\sigma(n)$, γ and λ. Early on in the translation process when $t < n/s$, no ribosomes have completed translation. As a result, in this time interval $u(t, 0) = \gamma/s$ (Figs. 2 and 3a when $t \leqslant 100$). When $2n/s > t \geqslant n/s$ the initial wave of ribosomes has completed recycling (Figs. 2 and 3a when $t = 200$). So in addition to the free ribosomes, recycled ribosomes are also contributing to $u(t, 0)$. Consequently $u(t, 0) = \gamma/s\,(1 + \lambda\,\sigma(n))$ for this range of t values. When $3n/s > t \geqslant 2n/s$, then $u(t, 0)$ is reinforced by a second wave of ribosome recycling (which was, in turn, reinforced by the first wave). Thus, $u(t, 0) = \gamma/s\,(1 + \lambda\,\sigma(n) + (\lambda\,\sigma(n))^2)$ (Fig. 3a when $t = 400$). If we define $\hat{u}(0)$ as the steady-state value of $u(t, 0)$, (16) indicates that,

$$\hat{u}(0) = \lim_{t \to \infty} u(t, 0) = \frac{\gamma}{s} \frac{1}{1 - \lambda\,\sigma(n)}. \tag{18}$$

Eq. (18) indicates that as long as $\lambda\,\sigma(n) < 1$, the steady-state is well defined.

With each successive wave of translation the initial amplitude of the wave, i.e. its value at $u(t, 0)$, increases due to recycling. However, as long as $\lambda < 1$ or $\sigma(n) < 1$, the size of each incremental increase due to reinforcement will always be smaller than the previous increase. As a result the initiation rate increases asymptotically towards $\hat{u}(0)$ (Fig. 4a).

By examining (16) more closely, we can assess the rate, in units of wave number r, at which $u(t, 0)$ approaches its asymptotic value. For example, if we

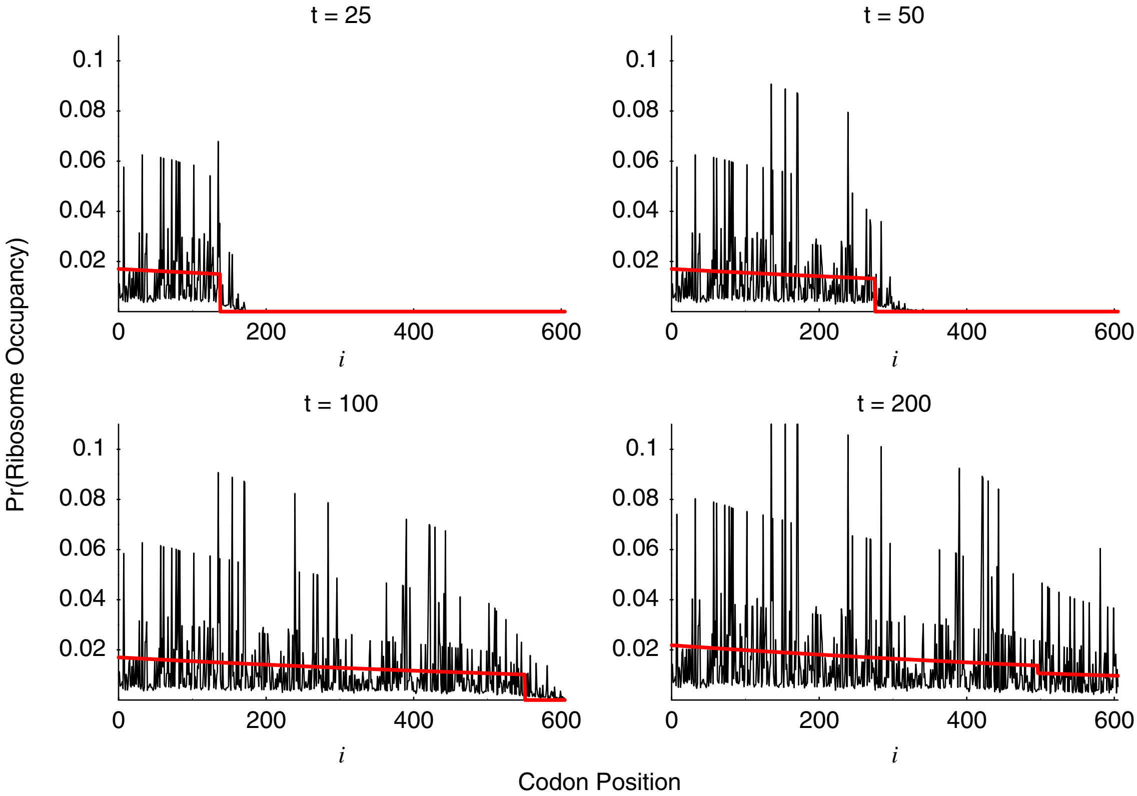

Fig. 2. Probability of ribosome occupancy, z_i, vs. codon position, i, at various time points for locus YOL158C. Both the discrete (–) and continuous (–) model versions are presented. Changes in the probability of a ribosome being found at a particular location can be viewed as a traveling wave which moves down the mRNA transcript. In the discrete model the rate of movement is heterogeneous due to differences in codon translation rates. At codons with low translation rates the probability spikes. In contrast, at codons with high translation rates the probability drops. In the continuous approximation the rate movement is assumed to be uniform, hence no spikes are observed. Once the wave front reaches the final codon it reinforces itself via ribosome recycling. As a result a higher second wave of ribosome occupancy probability moves down the mRNA transcript. The codon translation rates used are given in Table S1. The de novo ribosome initiation rate, γ was calculated from the locus' $\hat{k}$ value in Gilchrist and Wagner (in preparation) and with the ribosome recycling probability λ set to 0.5. The nonsense error rate was set to $b = 0.00515$.

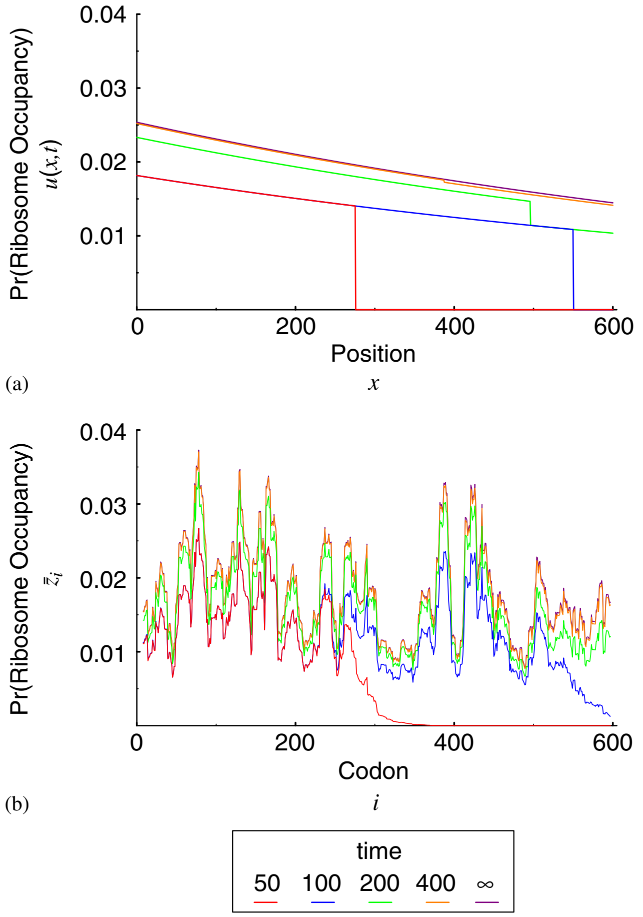

Fig. 3. Probability of ribosome occupancy vs. codon position for both the continuous model (a) the moving window mean value for the discrete model (b) at various time points. The mean values of the discrete model were calculated using a moving window of 15 codons (the approximate mRNA stretch covered by a eukaryotic ribosome) centered around i. The steady-state corresponds to $t = \infty$. The step-like changes in the distribution of ribosome probabilities in (a) represent new wave fronts of ribosome occupancy probabilities due to ribosome recycling. This behavior is less obvious but still apparent in (b). Note that each successive increase in ribosome occupancy probability is smaller than the previous increase. The solutions were calculated using the same locus and parameter values as in Fig. 2.

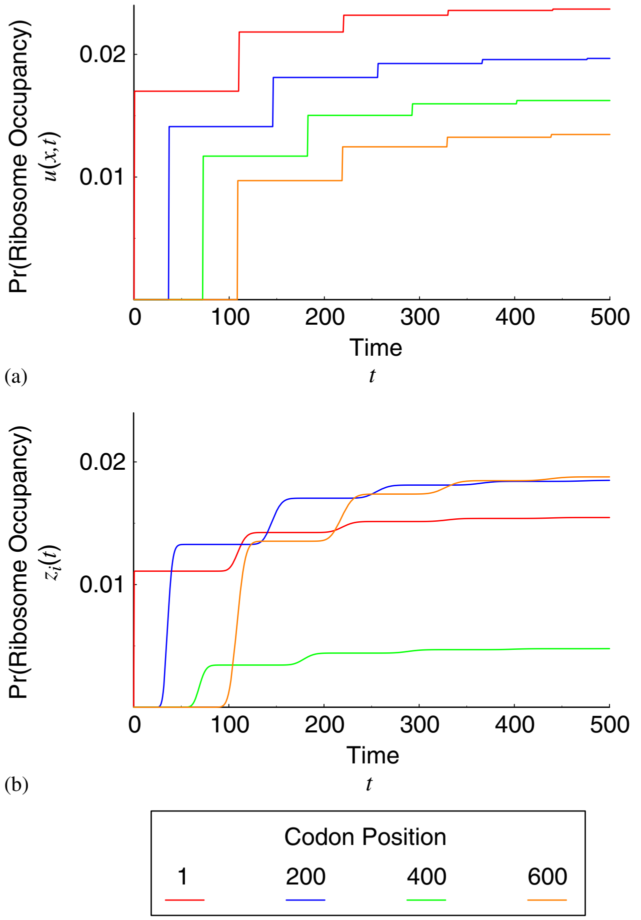

Fig. 4. Ribosome occupancy probability vs. time for codons: 1, 200, 400, and 600. (a) corresponds to the continuous model and (b) corresponds to the discrete model. (b) illustrates how heterogeneity in codon translation rates in the discrete codon model can lead to downstream codons having higher ribosome occupancy probabilities than upstream codons. Because the continuous approximation assumes a uniform translation rate, this type of behavior does not occur in (a). The step-like behavior of the curves reflects the shifts in ribosome initiation rates. The first step corresponds to the initial wave of new ribosomes binding to the mRNA transcript. Subsequent steps are due to ribosome recycling. Spacing between steps corresponds to the amount of time it takes to fully translate the mRNA transcript. The solutions were calculated using the same locus and parameter values as in Fig. 2.

define $r_{0.95}$ as the critical wave number which results in $u(t,0)$ being greater or equal to 0.95 times its steady-state value, $\hat{u}(0)$, it can be shown that,

$$r_{0.95} = \left\lfloor \frac{\ln(1-0.95)}{\ln(\lambda\,\sigma(n))} - 1 \right\rfloor. \tag{19}$$

Eq. (19) indicates that as the product of λ and $\sigma(n)$ increases, so does the number of waves it takes for $u(t,0)$ to closely approach its steady-state value, $\hat{u}(0)$ (Fig. 5). In the special case where $\lambda = 0$, $u(0,0) = \hat{u}(0)$. Thus, in the absence of ribosome recycling, the ribosome initiation rate is constant for all t. As a result, the system reaches the steady-state after the wave reaches the last codon.

To summarize, the solution to our partial differential equation indicates that the system can be viewed as a traveling wave propagating from the start codon propagating in the 5′ to 3′ direction. Because of nonsense errors, the amplitude of this wave decays as it moves towards the final codon. Both the speed at which the wave travels, s, and the rate at which it decays, $\sigma(n)^{x/n}$ are functions of the codon translation rates of an mRNA transcript, $\vec{c}$, and the nonsense error rate, b. The wave reinforces itself via ribosome recycling. The impact of this reinforcement is a function of the product of the ribosome recycling probability, λ, and the protein translation probability, $\sigma(n)$. The degree to which the

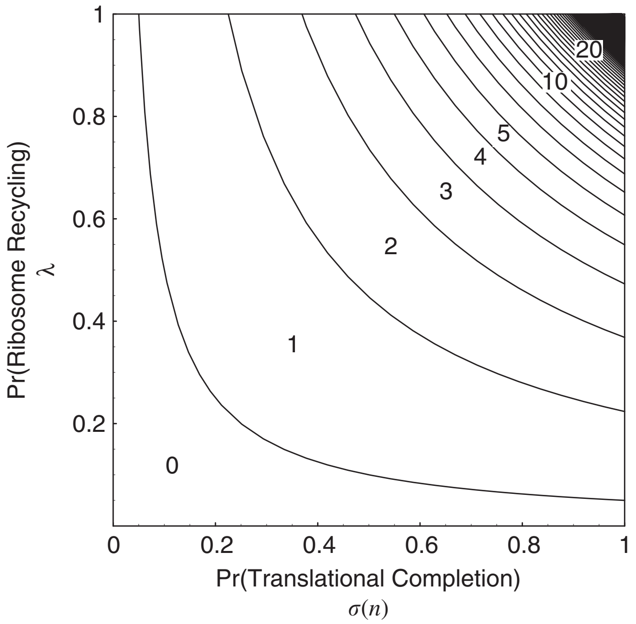

Fig. 5. Critical wave number, $r_{0.95}$, as a function of translational completion probability, $\sigma(n)$, and ribosome recycling probability, λ. The critical wave number $r_{0.95}$ represents the number of waves it takes for the ribosome density at the start codon, $u(t, 0)$, to be equal to 95% of its steady-state value, $\hat{u}(t)$. The function $r_{0.95}$ is an integer function and contour lines separate regions of different integer values. $r_{0.95}$ increases with the product of λ and $\sigma(n)$. See (19) for the exact solution.

wave reinforces itself decreases with each successive wave so that the system approaches a steady-state. The time it takes for the system to approach the steady-state, measured in terms of the number of waves r, increases with the product of λ and $\sigma(n)$.

3.6. Numerical solutions of the discrete model

We now compare our analytic solutions of the continuous codon model to numerical solutions of the discrete codon model. The most striking difference between the discrete and continuous solutions is the immense short-scale heterogeneity in the distribution of ribosome occupancy probabilities found in the discrete model (Fig. 2). Using the steady-state solution of the discrete model we will explicitly show that this short-scale heterogeneity in the ribosome occupancy probabilities, $\vec{z}(t)$ (the discrete counterpart of $u(t, x)$), results from heterogeneity of the codon translation rates in the mRNA. Firstly, this is because the probability that a ribosome will be found at a particular codon increases with the waiting time for the correct tRNA while at that codon. Secondly, this is because the expected waiting time is inversely proportional to a codon's translation rate. As a result ribosome densities spike at slowly translated codons and decrease at quickly translated codons. One consequence of this heterogeneity is that, unlike in the continuous case, the probability of a ribosome occurring at downstream codons is not always less than that of upstream codons (cf. Figs. 4a and b).

Although the continuous model cannot capture short-scale heterogeneity in ribosome occupancy probabilities, other aspects of its behavior are similar to the discrete model. For example, both models' solutions show a similar wave of ribosome occupancy probability moving along the transcript. (Although in the discrete model this wave front is less abrupt than in the continuous model and spreads out over time (Figs. 2 and 3b when $t \leqslant 100$).) In both the discrete and continuous models one also observes a general decline in the amplitude of this probability wave with increasing distance from the start codon (e.g. Fig. 2). Finally, in both models the wave reinforces itself through ribosome recycling (Fig. 2 when $t = 200$) and approach their steady-state in similar manners (Fig. 3).

3.7. Steady-state behavior

As our analysis of the discrete and continuous model dynamics indicates, as t becomes large the distribution of ribosome occupancy probabilities on a transcript, $\vec{z}$, asymptotically approaches its steady-state. We can solve for the steady-state of $\vec{z}$ by solving our discrete model under the condition that the derivatives in (1) are equal to zero. Using the accent ^ to denote the steady-state value of a variable, from (1) it is easy to show that,

$$\hat{z}_i = \frac{\hat{k}}{c_i + b}\sigma(i-1), \tag{20}$$

where

$$\hat{k} = \frac{\gamma}{(1 - \lambda\,\sigma(n))}. \tag{21}$$

Eq. (20) shows that the steady-state probability that a ribosome occupies codon i has a simple form. It is equal to the rate at which ribosomes initiate translation weighted by the probability that a ribosome will reach the codon, and divided by the total rate at which ribosomes leave the codon due to either translation or a nonsense errors. As mentioned earlier, the short-scale heterogeneity in $\hat{\vec{z}}$ is caused by the variation in expected waiting time between codons, where longer waiting times lead to higher densities. This is also evident from the steady-state solution as $\hat{z}_i$ is inversely proportional to $c_i + b$.

Fig. 6 presents typical forms of $\vec{z}$ at steady-state for a number of genes in the yeast genome. While the impact of heterogeneous c values is apparent in all cases, the impact of nonsense errors on the probability that a ribosome will reach a codon, $\sigma(i)$, is especially apparent in the longer genes.

From (20) we see that the steady-state values in $\hat{\vec{z}}$ are scaled by $\hat{k}$. Eq. (21) indicates that the steady-state initiation rate, $\hat{k}$, is an increasing function of the rate at

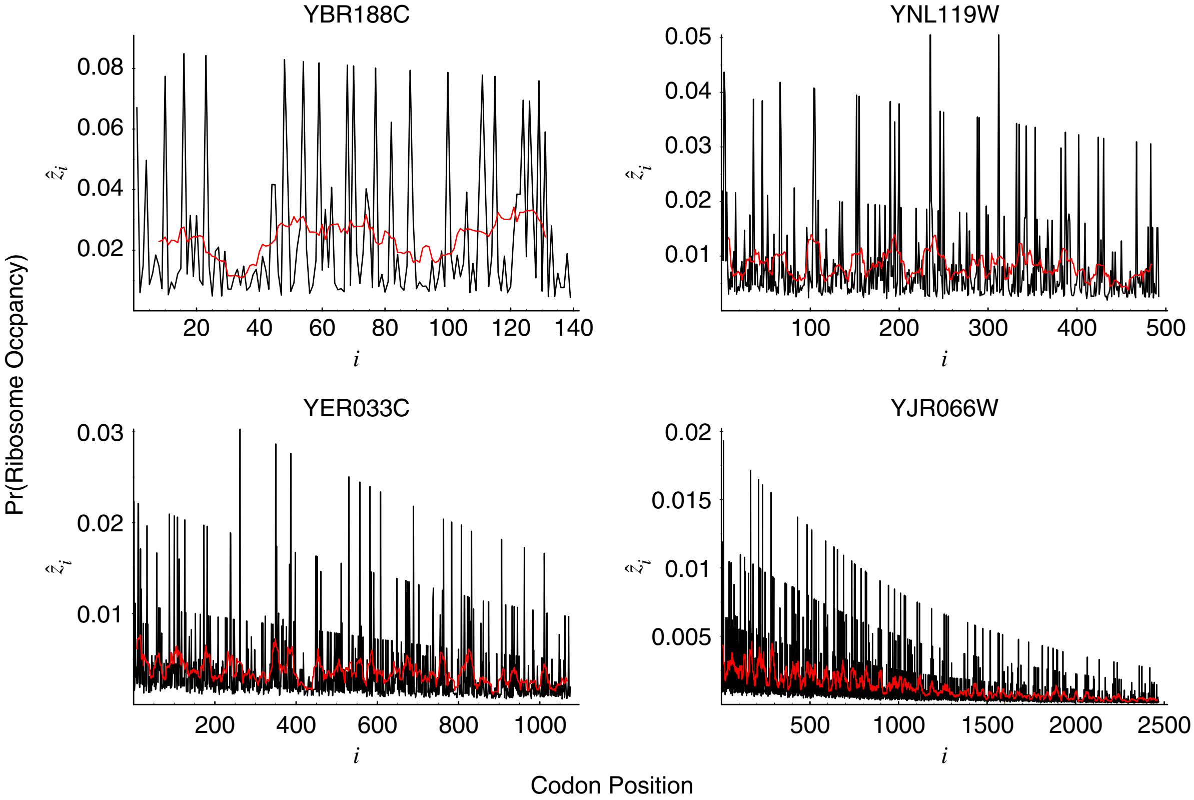

Fig. 6. Examples of ribosome occupancy probabilities, $\vec{z}$, at steady-state for four mRNA transcript of varying length. Although more apparent in the longer transcripts, the general decline in $\hat{z}_i$ is due to nonsense errors. (–) represents the mean value calculated using a moving window of width 15 codons. The de novo ribosome initiation rate, γ was calculated from the loci's $\hat{k}$ value in Gilchrist and Wagner (in preparation) and with the ribosome recycling probability λ set to 0.5. The nonsense error rate was set to $b = 0.00515$.

which new ribosomes initiate translation, γ, and the probability that a ribosome which begins translating the mRNA transcript will complete translation and be recycled, $\lambda\,\sigma(n)$. As $\lambda\,\sigma(n)$ approaches 1, $\hat{k}$ approaches infinity because once ribosomes begin translating a transcript they do not readily leave the translation-recycling loop. Note that (19) indicates that under this scenario the amount of time it takes to approach this state also approaches infinity.

From our definition of the steady-state ribosome initiation rate, $\hat{k}$, we can use (4) to get an explicit definition of the steady-state rate at which a protein becomes translated, $\hat{\tau}$,

$$\hat{\tau} = \frac{\gamma\,\sigma(n)}{(1 - \lambda\,\sigma(n))}. \qquad (22)$$

Note that in the absence of nonsense errors, $\hat{\tau} = \gamma/(1 - \lambda)$ and, consequently, the steady-state rate of protein synthesis is independent of codon usage.

The qualitative behavior of $\hat{\tau}$ is a function of the protein translation probability, $\sigma(n)$ and the ribosome recycling probability, λ. In contrast, the de novo ribosome initiation rate, γ, simply scales as $\hat{\tau}$. As both $\sigma(n)$ and λ are probabilities and are, therefore, bounded between 0 and 1, it is easy to evaluate the behavior of $\hat{\tau}$ (Fig. 7). That $\hat{\tau}$ must increase with λ is easy to see: λ essentially amplifies the effect of γ on $\hat{\tau}$. Because ribosome recycling only occurs with ribosomes that complete translation, the impact of λ on $\hat{\tau}$ is dependent on $\sigma(n)$. At low values of $\sigma(n)$, even a large value of λ has little impact on $\hat{\tau}$ because very few ribosomes can be recycled since very few complete translation. In contrast, when $\sigma(n)$ is close to 1, slight increases in λ lead to large increases in $\hat{\tau}$. This is because under this scenario, ribosomes which begin translating the protein are likely to complete translation. As a result, any increase in λ will increase the probability that a ribosome will translate the same mRNA transcript many more times.

4. The energetic cost of nonsense errors: ξ

4.1. Calculation and analysis

Based on the work above, we can now calculate the expected energetic cost of nonsense errors for each translational initiation event, ξ as defined in (5), for a

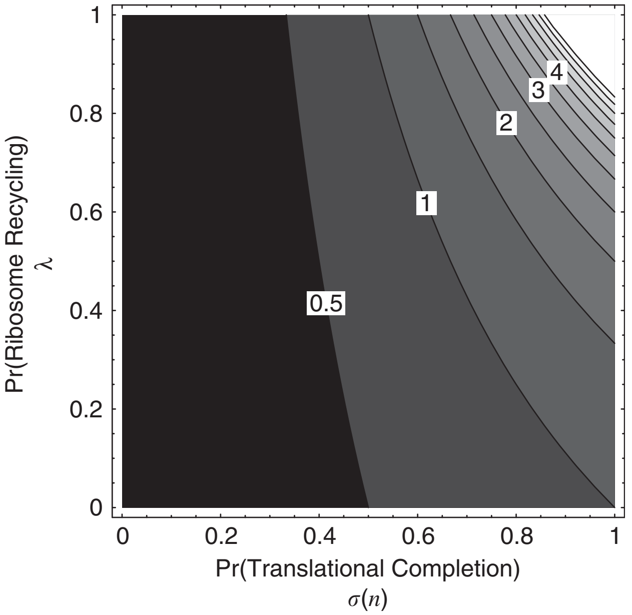

Fig. 7. Steady-state protein translation rate, $\hat{\tau}$, as a function of the ribosome recycling probability, λ, and the translational completion probability, $\sigma(n)$. The contours lines correspond to units of the de novo ribosome initiation rate, γ. Ribosome recycling can be viewed as an amplifying effect on the steady-state translation rate. The impact of ribosome recycling is strongly dependent on the protein translation probability $\sigma(n)$.

given transcript $\vec{c}$ as,

$$\xi(\vec{c}) = \sum_{i=1}^{n} (a_1 + a_2 i) \frac{b}{c_i} \sigma(i). \tag{23}$$

Having a fully defined definition of ξ allows us to analyze how ξ changes when we switch the codons c_k and c_m, where $k<m$. We use the symbol $\Delta\xi_{k,m}$ to represent the change in ξ caused by such a switch. From Eqs. (6) and (23) it can be shown that,

$$\Delta\xi_{k,m} = \left(1 - \frac{c_m}{c_m + b} \frac{c_k + b}{c_k}\right) \sum_{i=k}^{m-1} \sigma(k). \tag{24}$$

Eq. (24) implies that the sign of $\Delta\xi_{k,m}$, i.e., whether switching c_k and c_m increases or decreases the expected energetic cost of nonsense errors, is determined by the sign of the term in parentheses. Thus we find,

$$\Delta\xi_{k,m} = \begin{cases} >0, & c_m > c_k, \\ <0, & c_m < c_k. \end{cases} \tag{25}$$

Thus, the value of ξ increases if a faster codon late in the transcript is switched with a slower translating codon earlier in the transcript. This is 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. This result is consistent with verbal arguments put forth by (Eyre-Walker, 1996). Note that the parenthetical term in (24) is scaled by the sum of σ values affected by the codon switch. Since $\sigma(i)$ is always positive, this indicates that the impact of a switch increases with the distance between codons switched. Furthermore, because $\sigma(i)$ is also a monotonically decreasing function, switching two codons some distance apart at the start of a transcript would affect ξ more than switching two codons of the same distance towards the end of the transcript. This suggests that while the strength of selection on codon usage increases with codon position, the gradient of this selection force gradual decreases with position. In the natural world, mutations do not occur via codon switching. Nevertheless, the calculation of $\Delta\xi_{k,m}$ clarifies relationship between selection on codon usage and how the strength of selection on codon usage bias changes with codon position.

4.2. Comparing ξ to its predicted behavior

To test whether ξ and, indirectly our model, captures the underlying biological cost of nonsense errors we conducted the following computational experiment. For each confirmed protein transcript in the yeast genome, we generated a null distribution of 2500 transcripts where the codons used for each amino acid were randomly rearranged. For each transcript, we then calculated the average expected energetic cost of nonsense errors for the transcripts in the null set and then compared it to the observed cost expected cost of nonsense errors, $\xi(\vec{c})$. If the observed cost was less than the mean of the null set we scored that as a 1, indicating evidence for selection on codon order. If the observed cost was greater than the mean of the null set we scored that as a 0, indicating no evidence for selection on codon order.

If our model of protein translation and the expected nonsense error cost function based on it are invalid, then we would expect the observed costs of nonsense errors to be distributed in the same way as the values in the null set. On the other hand, if our model of protein translation and the expected nonsense error cost function based on it are generally valid *and* selection is acting on codon position as previously posited, then we would expect that observed costs of nonsense errors are generally less than the average cost in the null set.

We can test this idea explicitly by comparing the number of genes where the observed cost is less than the mean of their corresponding null set to the number expected based on a binomial model with a success rate of 0.5. Of the 5855 genes examined, 3562 (~61%) had an observed cost of nonsense errors less than the mean of the null distribution, a result which is highly significant ($p < 10^{-60}$). This result strongly supports the idea that selection on codon usage is site specific and increases with codon position.

Our cost function, ξ, describes the expected cost of nonsense errors for each initiation event of protein

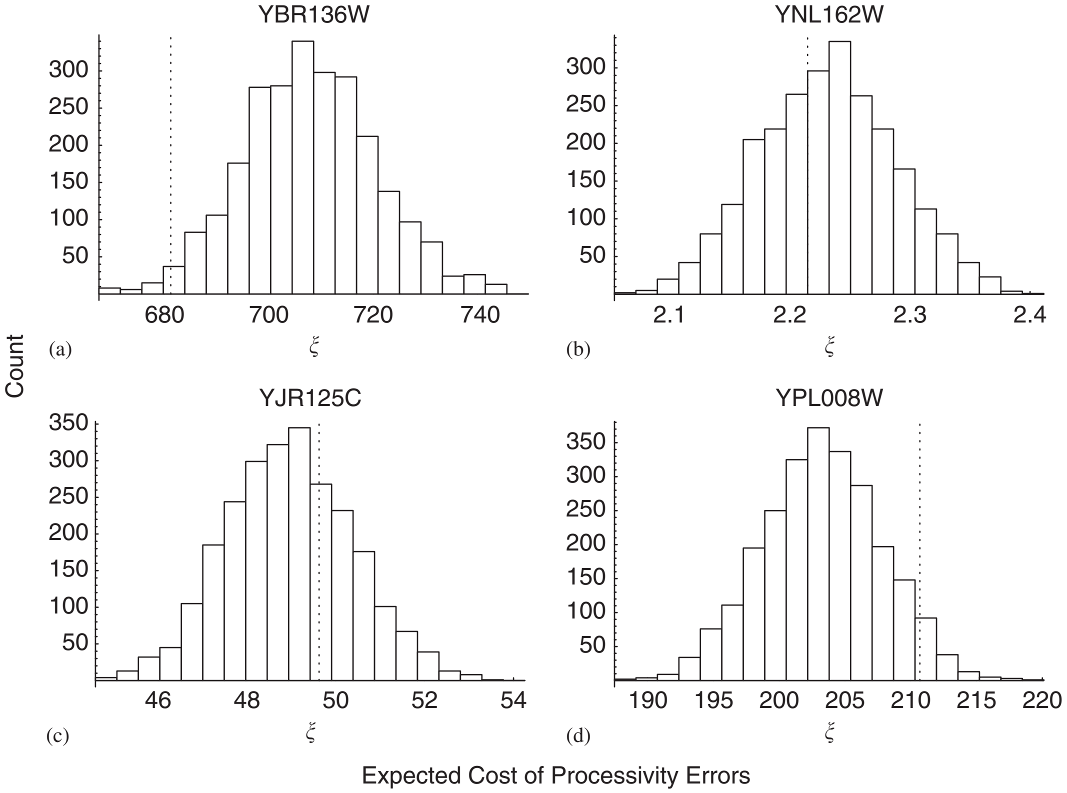

Fig. 8. Distribution of the expected cost of nonsense errors ξ for scrambled transcripts for four different loci. For each locus, 2500 scrambled transcripts were generated from the observed transcript by randomizing the order of codons for each amino acid separately. The observed expected cost of nonsense errors is indicated by the dotted line. The loci presented here were chosen to illustrate the range of behavior of these distributions and their corresponding observed values. In general the observed cost value are significantly less than the mean value of the scrambled cost values ($p < 10^{-60}$).

translation of a gene. As a result, we expect the strength of selection on genes to increase with their expression level. Consequently, we can validate our model further by asking whether the probability that a gene's observed cost of nonsense errors is less than expected increases with the expression level of the gene. In this analysis, we used the mRNA expression levels for yeast during exponential growth as presented in Beyer et al. (2004). These mRNA levels were derived from a wide collection of measurements from multiple laboratories and provide measurements for 5592 confirmed genes. The results of a logistic regression indicate that the probability of finding evidence of selection on codon order increases significantly with mRNA expression level ($p < 0.005$). More surprisingly, because the intercept of our regression is at 0.62, these results also indicate that the response to selection on codon order can even be detected in genes with weak expression levels (Figs. 8 and 9).

5. Discussion

A combination of a fine-grained model representing codons as discrete units and a coarse-grained continuous model of mRNAs shows that ribosomes translating any one mRNA can be thought of as a probability wave traveling through a heterogeneous medium. The wave undergoes decay caused by nonsense errors, but it also becomes reinforced via ribosome recycling. The heterogeneity of the medium (mRNA) is caused by the different rates at which individual codons are translated. Because of the heterogeneity in codon translation rates, the probability of a ribosome being found spikes at slow codons and drops at fast codons. If nonsense errors occur at a constant rate per unit time, then they are more likely to occur at slow codons than at fast codons. Therefore, codon usage bias not only causes short-scale heterogeneity in the ribosome occupancy probabilities along an mRNA, but it also affects the probability that a nonsense error will occur along an mRNA. It is likely that the rate of nonsense errors varies between codons, although we did not consider the possibility in our model. For example, Freistroffer et al. (2000) show that near-cognates of the stop codons are more often mistranslated by the release factors than other codons. However, the authors also conclude that only a "small proportion of nonsense failures can be attributed" to this process. As a result, incorporating such information

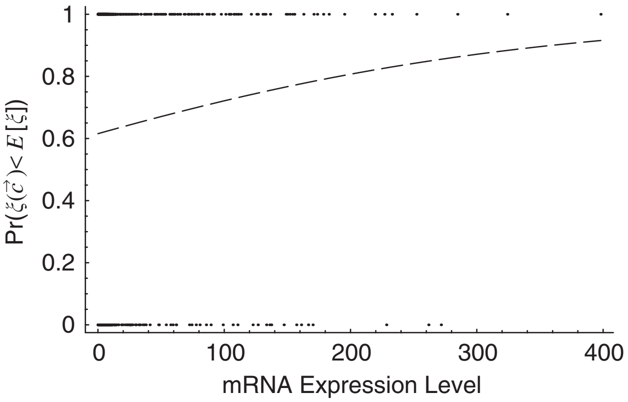

Fig. 9. Distribution of genes with evidence for selection on codon order vs. mRNA expression level data from Beyer et al. (2004). Genes were considered to have evidence for selection when their expected cost of nonsense errors of the transcript, $\xi(\vec{c})$, was less than the mean expected cost, $E(\xi)$ from a set of 2500 transcripts where codon order for each amino acid was randomized. The dashed line (— —) represents the best-fit line based on a logistic regression ($N = 5562$, $p < 0.005$) with parameters $\alpha = 0.47$ and $\beta = 0.0048$ (Agresti, 2002). The results indicate that the strength of selection on codon usage increases with codon position and mRNA expression level.

will be useful in future studies, but we do not expect this added level of model complexity to fundamentally change our findings.

There are numerous explicit models of protein translation in the literature (Bergmann and Lodish, 1979; Harley et al., 1981; Menninger, 1983; Liljenström and von Heijne, 1987; Bulmer, 1991; Zhang et al., 1994; Chou, 2003). Most of these models focus on the protein production rate of the entire cell and tend to ignore either nonsense errors and/or ribosome recycling. One common complication often introduced in these models is inter-ribosomal interference (i.e. the blocking of the translation of one ribosome by another ribosome immediately 3′ from it). Historically, introducing this complication generally precluded any analytic analysis, causing researchers to resort to simulation (Bergmann and Lodish, 1979; Zhang et al., 1994). A notable exception is a recent study by Chou (2003) which presents an analytic model of protein translation that included such interference as well as ribosome recycling. While this work is elegant and clearly a major step forward, the approach ignores nonsense errors and heterogeneity in codon translation rates.

The existing studies of inter-ribosomal interference indicate that such interference can reduce the translation rate of a protein and results from either the clumping of slowly translating codons or from high initiation rates. If inter-ribosomal interference does occur, it is unclear how it will ultimately affect the rate of nonsense errors. On one hand, such interference is likely to increase the probability of drop-off errors during translation. On the other hand, interference could possibly decrease the rate of frameshift errors if ribosome sizes are such that ribosomes in contact with one another are properly spaced (i.e., the distance between ribosomes, when measured in nucleotides, is a multiple of three). Protein initiation rates are largely unknown, but the data from Arava et al. (2003) shows that ribosome densities on mRNA transcripts are generally quite low compared to their possible saturation values. This evidence suggests that inter-ribosomal interference is likely to be rare in yeast.

Despite the large literature produced over the last 25 years on codon usage bias, only a few notable studies have focused on the role nonsense errors may play in shaping codon usage bias (Eyre-Walker, 1996; Berg and Silva, 1997; Hooper and Berg, 2000; Qin et al., 2004). This is surprising given that prematurely terminated and nonfunctional protein products that result from nonsense errors appear to be quite common (Manley, 1978; Tsung et al., 1989; Jorgensen and Kurland, 1990; Kurland, 1992). Nonsense errors are also likely to be quite costly since they can tie up essential cell resource such as tRNA molecules (Dincbas et al., 1999; Cruz-Vera et al., 2004), interfere with other cellular processes (Menninger, 1978), and waste many high-energy phosphate bonds.

In our study we used 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. We indirectly test this cost function and the underlying translation model on which it is based by asking whether or not we are able to detect a response to an increasing selection gradient on codon usage bias as predicted by Eyre-Walker (1996).

Gradients in codon usage bias have been previously investigated in a number of studies (Liljenström and von Heijne, 1987; Bulmer, 1988b; Eyre-Walker and Bulmer, 1993; Berg and Silva, 1997; Hooper and Berg, 2000; Qin et al., 2004). Most of these studies have focused on *E. coli* and the results of the earlier work were hampered by small sample sizes. The recent study by Qin et al. (2004) is the most statistically sophisticated one to date and utilized entire genomes for their analysis. These researchers applied a special regression technique to the *Drosophila melanogaster*, yeast and a number of prokaryote genomes. In all but *D. melanogaster*, Qin et al. (2004) detected consistent increases in codon usage bias with codon position. They also found that further increases in the gradients were commonly noticeable towards the end of the transcripts, which is consistent with our analysis of the effect of codon switching on the expected energetic cost function. In fact, in a number of prokaryotes codon bias actually dropped towards the very end of the transcript. If nonsense errors which occur late in a transcript result in fully or partially functional proteins, then selection on these final codons will be relaxed and this drop in codon usage bias can be explained. Given that gradients in codon usage occur, we view our detection of a response to position dependent selection on codon usage bias using our cost function and mRNA expression levels as indirect validation of this function and the protein translation model on which it is based.

In this study, much of our focus has been on nonsense errors and their costs. However, most discussions of selection on codon bias focus on selection for increased translational efficiency. Translational efficiency is generally thought of as the rate at which the ribosome population within a cell is forming peptide bonds relative to some hypothetical maximum rate. Interestingly, studies of translational efficiency generally ignore the impact of nonsense errors. Yet, if one assumes that incomplete peptides resulting from nonsense errors are largely non-functional, then it is clear that nonsense errors also affect the overall translational efficiency of the ribosome population. This is because nonsense errors decrease the number of peptide bonds created by the ribosome population that actually end up in a functional protein.

Ribosome recycling is likely to have evolved as a means of increasing the translational efficiency of the ribosome population. This is because ribosome recycling decreases the amount of time a ribosome spends between completing translation and initiation. However, ribosome recycling is only likely to occur if the ribosome translates the entire transcript. Thus we conclude that the overall efficacy of ribosome recycling is limited by the translational probability of a transcript which is determined, in turn, by the probability of a nonsense error at any one of the codons. As a result, we find that translational efficiency and nonsense errors are inextricably tied together and we are left with a clearer understanding of how codon usage bias underlies this interaction.

Acknowledgements

We thank NIH for support through Grant GM63882. AW would like to thank the Institut des Hautes Etudes for sabbatical support. MG would like to thank Aaron King for his helpful discussions. Finally, we also thank two anonymous reviewers for their constructive comments and suggestions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2005.08.007.

References

Agresti, A., 2002. Categorical Data Analysis. Wiley Series in Probability and Statistics, second ed. Wiley, Hoboken, NJ.

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

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, Y.L., Storey, J.D., Liu, C.L., Brown, P.O., Herschlag, D., 2003. Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. Proc. Natl Acad. Sci. U.S.A. 100, 3889–3894.

Bennetzen, J.L., Hall, B.D., 1982. Codon selection in yeast. J. Biol. Chem. 257, 3026–3031.

Berg, O.G., Silva, P.J.N., 1997. Codon bias in *Escherichia coli*: the influence of codon context on mutation and selection. Nucleic Acids Res. 25, 1397–1404.

Bergmann, J.E., Lodish, H.F., 1979. Kinetic-model of protein-synthesis—application to hemoglobin-synthesis and translational control. J. Biol. Chem. 254, 1927–1937.

Bernardi, G., Bernardi, G., 1986. Compositional constraints and genome evolution. J. Mol. Evol. 24, 1–11.

Beyer, A., Hollunder, J., Nasheuer, H.-P., Wilhelm, T., 2004. Post-transcriptional expression regulation in the yeast *Saccharomyces cerevisiae* on a genomic scale. Mol. Cell. Proteomics 3 (11), 1083–1092.

Birdsell, J.A., 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol. Biol. Evol. 19, 1181–1197.

Bulmer, M., 1988a. Are codon usage patterns in unicellular organisms determined by selection–mutation balance. J. Evol. Biol. 1, 15–26.

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.

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

Chavancy, G., Garel, J.P., 1981. Does quantitative transfer-RNA adaptation to codon content in messenger-RNA optimize the ribosomal translation efficiency—proposal for a translation system model. Biochimie 63, 187–195.

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

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

Cruz-Vera, L.R., Magos-Castro, M.A., Zamora-Romo, E., Guarneros, G., 2004. Ribosome stalling and peptidyl-tRNA drop-off during translational delay at aga codons. Nucleic Acids Res. 32, 4462–4468.

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

Dincbas, V., Heurgue-Hamard, V., Buckingham, R.H., Karimi, R., Ehrenberg, M., 1999. Shutdown in protein synthesis due to the expression of mini-genes in bacteria. J. Mol. Biol. 291, 745–759.

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.

Eyre-Walker, A., Bulmer, M., 1993. Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res. 21, 4599–4603.

Freistroffer, D.V., Kwiatkowski, M., Buckingham, R.H., Ehrenberg, M., 2000. The accuracy of codon recognition by polypeptide release factors. Proc. Natl Acad. Sci. USA 97, 2046–2051.

Gallie, D.R., 1991. The cap and poly(A) tail function synergistically to regulate messenger-RNA translational efficiency. Genes Dev. 5, 2108–2116.

Ghosh, T.C., Gupta, S.K., Majumdar, S., 2000. Studies on codon usage in *Entamoeba histolytica*. Int. J. Parasitol. 30, 715–722.

Gilchrist, M.A., Salter, L.A., Wagner, A., 2004. A statistical framework for combining and interpreting proteomic datasets. Bioinform. 20, 689–700.

Gouy, M., Grantham, R., 1980. Polypeptide elongation and transfer-RNA cycling in *Escherichia coli*—a dynamic approach. FEBS Lett. 115, 151–155.

Harley, C.B., Pollard, J.W., Stanners, C.P., Goldstein, S., 1981. Model for messenger-RNA translation during amino-acid starvation applied to the calculation of protein synthetic error rates. J. Biol. Chem. 256, 786–794.

Hooper, S.D., Berg, O.G., 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., 1982. Correlation between the abundance of yeast transfer RNAs and the occurance of the respective codons in protein genes. J. Mol. Biol. 158, 573–597.

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

Jacobson, A., 1996. Poly(A) metabolism and translation: The closed-loop model. In: Hershey, J.W., Mathews, M.B., Sonenberg, N. (Eds.), Translational Control. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 451–480.

Jansen, R., Yu, H.Y., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S.B., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M., 2003. A bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302, 449–453.

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

Kapp, L.D., Lorsch, J.R., 2004. The molecular mechanics of eukaryotic translation. Annu. Rev. Biochem. 73, 657–704.

Khaleghpour, K., Svitkin, Y.V., Craig, A.W., DeMaria, C.T., Deo, R.C., Burley, S.K., Sonenberg, N., 2001. Translational repression by a novel partner of human poly(A) binding protein. PAIP2. Mol. Cell 7, 205–216.

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.

Kozak, M., 2002. Pushing the limits of the scanning mechanism for initiation of translation. Gene 299, 1–34.

Kurland, C.G., 1992. Translational accuracy and the fitness of bacteria. Annu. Rev. Genet. 26, 29–50.

Lee, I., Date, S.V., Adai, A.T., Marcotte, E.M., 2004. A probabilistic functional network of yeast genes. Science 306, 1555–1558.

Liljenström, H., von Heijne, G., 1987. Translation rate modification by preferential codon usage: intragenic position effects. J. Theor. Biol. 124, 43–55.

Manley, J.L., 1978. Synthesis and degradation of termination and premature-termination fragments of beta-galactosidase invitro and invivo. J. Mol. Biol. 125, 407–432.

McVean, G.A.T., Charlesworth, B., 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genet. Res. 74, 145–158.

Menninger, J.R., 1978. Accumulation as peptidyl-transfer RNA of isoaccepting transfer-RNA families in *Escherichia coli* with temperature-sensitive peptidyl-transfer RNA hydrolase. J. Biol. Chem. 253, 6808–6813.

Menninger, J.R., 1983. Computer-simulation of ribosome editing. J. Mol. Biol. 171, 383–399.

Mougel, F., Manichanh, C., N'Guyen, G.D., Termier, M., 2004. Genomic choice of codons in 16 microbial species. J. Biomol. Struct. Dynam. 22, 315–329.

Murray, J.D., 1993. Mathematical Biology, second ed. Springer, Berlin.

Musto, H., Romero, H., Zavala, A., Jabbari, K., Bernardi, G., 1999. Synonymous codon choices in the extremely GC-poor genome of *Plasmodium falciparum*: compositional constraints and translational selection. J. Mol. Evol. 49, 27–35.

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.

Niepel, M., Ling, J., Gallie, D.R., 1999. Secondary structure in the 5′-leader or 3′-untranslated region reduces protein yield but does not affect the functional interaction between the 5′-cap and the poly(A) tail. FEBS Lett. 462, 79–84.

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

Qin, H., Wu, W.B., Comeron, J.M., Kreitman, M., Li, W.H., 2004. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. Genetics 168, 2245–2260.

Rajkowitsch, L., Vilela, C., Berthelot, K., Ramirez, C.V., McCarthy, J.E.G., 2004. Reinitiation and recycling are distinct processes occurring downstream of translation termination in yeast. J. Mol. Biol. 335, 71–85.

Sachs, A., 2000. Physical and functional interactions between the mRNA cap structure and the poly(A) tail. In: Sonenberg, N., Hershey, J., Mathews, M. (Eds.), Translational Control of Gene Expression. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 447–465.

Sharp, P.M., Li, W.H., 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.

Sorensen, M.A., Kurland, C.G., Pedersen, S., 1989. Codon usage determines translation rate in *Escherichia coli*. J. Mol. Biol. 207, 365–377.

Thomas, L.K., Dix, D.B., Thompson, R.C., 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.

Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B., Botstein, D., 2003. A bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). Proc. Natl Acad. Sci. USA. 100, 8348–8353.

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.

Tuite, M.F., 1989. Protein synthesis. In: Rose, A.H., Harrison, J.S. (Eds.), The Yeasts, vol. 3, second ed. Academic Press Ltd., New York, pp. 161–204.

Varenne, S., Buc, J., Lloubes, R., Lazdunski, C., 1984. Translation is a non-uniform process—effect of transfer-RNA availability on the rate of elongation of nascent polypeptide-chains. J. Mol. Biol. 180, 549–576.

Wagner, A., 2000. Inferring lifestyle from gene expression patterns. Mol. Biol. Evol. 17, 1985–1987.

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

Welch, E., Wang, W., Peltz, S., 2000. Translational termination: It's not the end of the story. In: Sonenberg, N., Hershey, J., Mathews, M. (Eds.), Translational Control of Gene Expression. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 467–485.

Wolfram Research Inc., 2003. Mathematica, version 5.0. Wolfram Research Inc., Champaign, IL.

Xia, X.H., 1996. Maximizing transcription efficiency causes codon usage bias. Genetics 144, 1309–1320.

Xia, X.H., 1998. How optimized is the translational machinery in *Escherichia coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae*? Genetics 149, 37–44.

Zhang, S.P., Goldman, E., Zubay, G., 1994. Clustering of low usage codons and ribosome movement. J. Theor. Biol. 170, 339–354.