8.592J Final Project: A model of codon usage and ribosome traffic jams

Jue Wang

May 15, 2011

Background

Translation of messenger RNA (mRNA) into protein is a central biological process that is under constant evolutionary pressure to be as fast and accurate as possible. Under typical conditions, both the rate (Ikemura, 1981) and accuracy (Akashi, 1994) of translation are determined by the availability of transfer RNAs (tRNAs) cognate to (i.e. matching) the codons of a gene being translated. Because most amino acids can be encoded by multiple codons, and different codons are generally not read by equal concentrations of cognate tRNAs, even synonymous mutations can affect the fitness of an organism by altering the production rate and quality of its proteins (Drummond and Wilke, 2008). Translation optimality thus presents a selective constraint on the primary sequence of genes beyond protein sequence. For this project I will consider a specific codon usage problem that is thought to arise in the translation of high expression mRNAs.

Translation roughly proceeds by the following sequence of steps (Uemura et al., 2010): a ribosome binds to an mRNA just upstream of the coding sequence for a gene; it then "jumps" from one codon to the next at a rate limited by binding of cognate tRNA; after the last codon has been translated, the ribosome dissociates from the transcript and releases a fully translated protein. For most genes, translation initiation is slow and therefore sets the overall protein production rate. However, for certain genes in fast-growing cells, initiation is sufficiently frequent that ribosomes can accumulate on an mRNA. In this regime, the steady state flux of ribosomes off of the transcript may still be determined by the initiation rate, but the average ribosome density on a transcript will vary widely according to the physical arrangement of codons. For example, an inefficient "bottleneck" codon in the middle of a transcript may slow translation and cause ribosome "traffic jams" even if the codons upstream are all efficient (Navon and Pilpel, 2011). Because ribosomes are limited in number and expensive to produce, cells are under selective pressure to minimize the number of ribosomes sequestered on transcripts for a given rate of protein production. Indeed, a recent study found that in high-expression transcripts from several model organisms, there is a region at the 5' end where low-efficiency codons are preferred (Tuller et al., 2010). The authors argued that this "throttle" region can help a cell to reduce ribosome density on transcripts and thus increase the efficiency with which it uses ribosomes.

I will explore the behavior of codon "throttle" regions using a simplified model of translation, and look at how ribosome traffic jams on transcripts are affected by initiation rates and the spatial sequence of codon jump rates. In particular I will identify a tradeoff between protein production rate and ribosome efficiency that I think is an important contributor to the fitness landscape on which cells are tuning their codon usage.

Model

We represent an mRNA transcript as a one-dimensional lattice of length N, where each lattice site corresponds to a codon. We treat the ribosome as a particle that moves in one direction along the lattice and occupies exactly one site at any given time. Actual ribosomes in, for example, yeast, have a ribosome "footprint" of about15 codons (Ingolia et al., 2009), but we ignore this for simplicity, and because it is unlikely to affect the qualitative behavior of the model.

We can give the model discrete-time dynamics as follows. On every time step, randomly pick one of $N+1$ transitions with uniform probability: initiate a new ribosome or move an existing ribosome at one of the N sites. If initiation is chosen and site 1 is empty, a ribosome appears on site 1 with probability α. Do nothing if site 1 is occupied or with probability $1-\alpha$ if it is empty. If one of the sites $\{i;\ 1 \leq i \leq N\}$

is chosen for a jump, and if site i is occupied and site $i+1$ is empty, the particle at site i moves to $i+1$ with jump probability γ_i. In the special case that $i = N$, a jump transition takes the ribosome off the lattice; this signifies termination of translation and occurs with termination probability β. Once a transition has been performed, repeat again.

$$\xrightarrow{\alpha} 1 \xrightarrow{\gamma_1} 2 \xrightarrow{\gamma_2} \dots \xrightarrow{\gamma_{N-2}} N-1 \xrightarrow{\gamma_{N-1}} N \xrightarrow{\beta}$$

This model is known as the Totally Asymmetric Simple Exclusion Process (TASEP) and belongs to a general class of asymmetric exclusion processes that have been studied extensively in the statistical physics literature. In particular, for a TASEP with uniform jump probabilities, i.e. $\gamma_i = 1$ for $1 \leq i \leq N$, it is possible to calculate analytically the average steady-state occupancies at each site, revealing a surprisingly rich phase behavior as α and β are varied (Derrida et al., 1992; Schütz and Domany, 1993). Other studies have focused on the effect of including one slow codon on a backdrop of otherwise uniformly fast codons (Janowsky and Lebowitz, 1994), or the effect of varying the footprint size of the particle (Lakatos and Chou, 2003). TASEPs have also been extended to incorporate arbitrary jump rates associated either with lattice positions or particles– the former is referred to as quenched disorder (Krug, 2000), and is the form of the model most relevant to codon efficiency in translation.

The outputs of the model we will examine are the average steady-state total occupancy θ (ribosome sequestration) and the average steady state termination flux J (protein production) for different configurations of jump rates. An analytical calculation of these quantities is possible (Kolwankar and Punnoose, 2000) but difficult, so we will obtain them by stochastic simulation using the Gillespie (or Monte Carlo) algorithm (Gillespie). This requires adapting the above scheme to continuous time as follows. Instead of transition probabilities we have initiation, jump, and termination rates of α, γ_i, and β, which are proportional to the respective discrete-time probabilities. Each individual transition occurs with an exponentially distributed waiting time, which means that an ensemble of transitions with rates $\{r_i;\ 1 \leq i \leq k\}$ and combined rate $r_{tot} = \sum_i r_i$ has a minimum waiting time Δt_{min} that is also exponentially distributed, as $p(\Delta t_{min}) = r_{tot} e^{-r_{tot}\Delta t_{min}}$. In each simulation step we begin by taking account of all accessible transitions (i.e. excluding blocked ribosomes), compute the waiting time until a generic transition, choosing a particular transition from the accessible transitions with a probability proportional to its rate, and updating the simulation time and state accordingly. No computational time is spent on futile transitions (such as a blocked initiation or jump event) as these are accounted for in the waiting times.

Simulations were performed in Matlab on a Lenovo Thinkpad T61 PC with an Intel Core Duo 2.2 GHz processor. Unless otherwise noted, simulations were initialized with an empty lattice of size $N = 200$ and run for 500000 MCS (Monte Carlo Steps). Total occupancy reached apparent steady state after less than 100000 MCS; therefore, the transient from MCS 1 to MCS 100000 was omitted when computing steady state averages.

Results

First let's consider the dependence of occupancy θ and flux J on the parameters α and β for uniform jump rates $\gamma_i = \gamma$. Intuitively, we would expect that when $\alpha \ll \gamma$, $\theta \approx 0$ and $J \approx \alpha$; this is equivalent to saying that when initiation is slow relative to elongation, transcripts tend to be translated by one ribosome at a time, and protein production rate is set by the initiation rate. On the other hand, if $\beta \ll \gamma \approx \alpha$, then $J \approx \beta$ and $\theta \approx N$ because there is a "bottleneck" at the end of the transcript and ribosomes tend to cover the entire mRNA. If both initiation and termination are fast, i.e. $\alpha \approx \beta \gg \gamma$, then $\theta > 0$ and $J < \alpha$ because ribosome traffic jams will develop, and flux can't keep up with initiation rate because the ribosome binding site will be occluded some fraction of the time. Indeed, these parameter regimes represent well known features of the TASEP model, and the phase transitions between them have been analyzed elsewhere in detail for $\gamma_i = 1$ (Schütz and Domany, 1993); here I demonstrate these phase transitions by simulations (Fig. 1a) as a verification of my code and as a reference point. Indeed, we see low- and high-occupancy regimes as well as an intermediate-occupancy, high-flux regime. Although not obvious in the heat map plots, flux in the low-flux regime is linear in α and β, as we also expected.

This qualitative picture of the phase landscape does not change when we introduce a stretch of 10 slow codons ($\gamma_i = 0.5$) at the beginning of the lattice (Fig 1b). However, the occupancy and the maximal flux in the high-flux regime are decreased. In a sense, the slow codon region serves to "throttle down" initiation so that there is a low *effective* initiation rate, which results in a low density of ribosomes on the rest of the transcript.

How might a low ribosome density benefit a cell? If we consider the flux per ribosome J/θ (Fig 1a,b, third column), this gives us a measure of how effi-

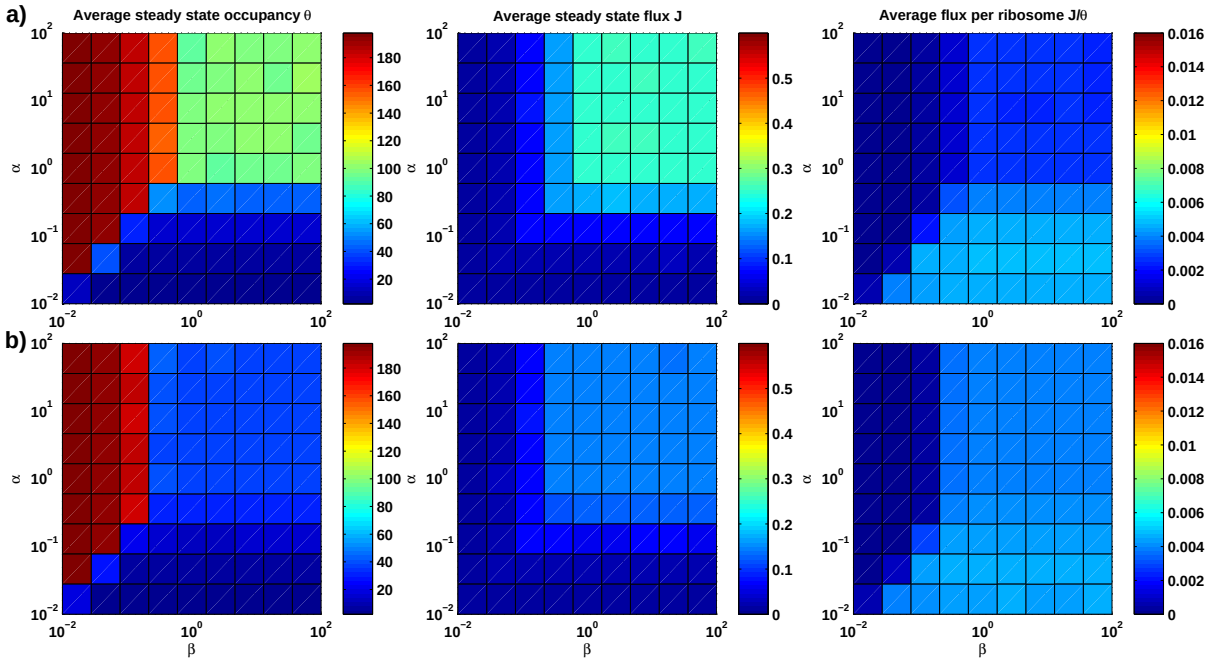


Figure 1: Phase plots for total lattice occupancy θ, flux J, and flux per ribosome J/θ with varying α and β and $N = 200$. All values computed as steady state averages for (a) uniform jump rates, i.e. all $\gamma_i = 1$ and (b) in the presence of a low-efficiency "step" at the beginning of the lattice, i.e. $\gamma_i = 0.5$ for $1 \leq i \leq 10$ and $\gamma_i = 1$ otherwise. For both jump-rate configurations, occupancy transitions sharply between low ($\theta \approx 0$), half-maximal ($\theta \approx N/2$, a) or intermediate ($0 < \theta < N/2$, b), and maximal ($\theta \approx N$) regimes. Flux transitions between linear and saturated maximal regimes, where $J \approx 0.25$ for the uniform case (a) and $J \approx 0.15$ when a step is added (b). For the uniform case (a), the critical values of the parameters have been calculated as $\alpha_c = 0.5$ and $\beta_c = 0.5$.

ciently each ribosome is being used to make protein. Clearly, slowing down initiation (either by decreasing α or adding a codon "throttle") will improve ribosome efficiency by decreasing θ, but this comes at a cost of decreased J, which we can think of as how efficiently each mRNA is being used. This tradeoff, between efficiency with respect to ribosomes and efficiency with respect to mRNA, can be seen by plotting J/θ versus J (Fig. 2) and comparing the cases of a "flat" codon profile and one in which a throttle or "step" exists. In both cases, increasing the value of α decreases J/θ and increases J, and at high α the value of J/θ is higher for the "step" transcript. Thus, if for some reason initiation is fixed at a high level in a cell, but a protein isn't needed in maximal quantities, then a codon throttle would reduce ribosome sequestration and benefit the cell. On the other hand, there is no obvious reason why the value of α would need to be fixed in general–indeed, because initiation is in reality extremely complex and involves many aspects of the ribosome and transcript (milon 2008), there should be in principle many ways for the cell to tune α over several orders of magnitude. The shape of the two curves also indicates that given a "step" transcript and some α, one can always find a different value of α for which the "flat" transcript performs better in both J/θ and J. In other words, in the comparison we've set up here, there is no reason to favor a codon "throttle" if initiation can be tuned freely.

Of course, this comparison is in some sense unfair, because we are comparing a transcript with all maximally efficient codons with a transcript that has some less efficient codons. To see the effect of spatial positioning of codons without changing codon usage, let's instead compare transcripts which simply contain different orderings of the same set of randomly selected jump rates. We will define a "random" transcript as one in which the γ_i are drawn from a Gaussian distribution with mean 0.75 and standard deviation 0.1. After drawing we enforce boundary conditions $\gamma_i \geq 0.01$ and $\gamma_i \leq 1$ by replacing any out-of-bounds values with the respective boundary values, which represent the slowest and fastest codons available in the cell. Associated with a given "random" transcript we will design a "throttled" transcript which contains the same γ_i but with the lowest 10 γ_i are identified sequentially and swapped into the first respective 10 positions. Furthermore we will also define a "ramp" transcript that is identical to the throttled one except after swapping, the 10 codons are also rearranged in ascending order so that the slowest codon on the entire transcript is first.

The results of this comparison are shown in Fig. 3. As in above comparison between a "step" and "flat"

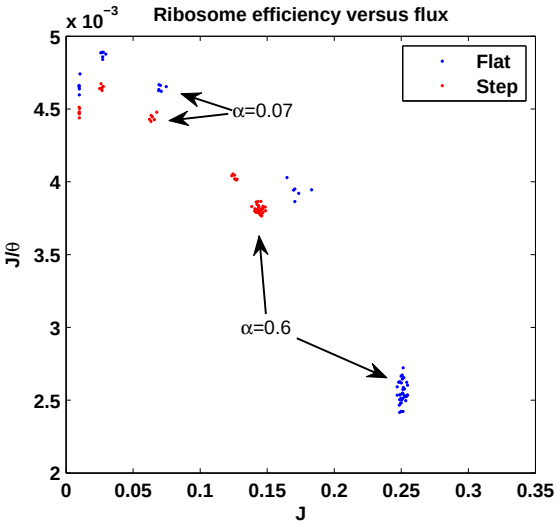


Figure 2: Tradeoff between ribosome efficiency J/θ and mRNA efficiency J. Simulations are the same as those used in Fig. 1. Data where $\beta < \beta_c$ has been omitted to consider only the transition between low-density and high-flux states. The codon "step" provides an improvement in J/θ at fixed large α, but not when α is small. Furthermore, because the "flat" transcript data points lie on a curve further from the origin than the points for the "step" transcript, we know that for any value of α in the "step" there exists a value of α on the "flat" transcript that yields both higher J and J/θ.

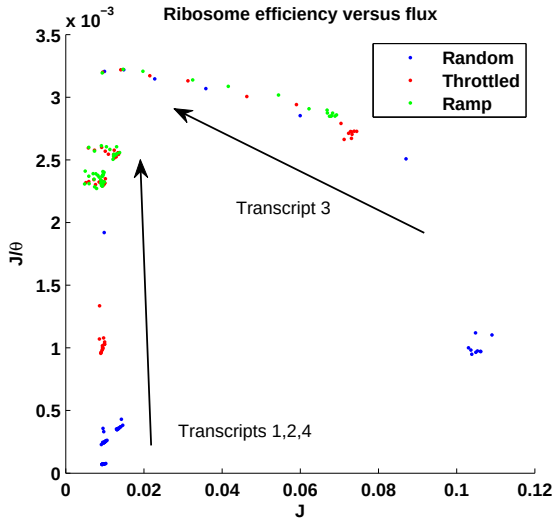
transcript, the "throttled" and "ramp" transcripts exhibit much higher ribosomal efficiency J/θ than the random transcript. Furthermore, for all the transcripts, increasing J/θ comes at a cost of decreasing J, although one of the transcripts has an inherently higher J due to its more efficient codon composition (Fig. 3b). Interestingly, there is no obvious difference in optimality between the throttled and ramp forms of 3 of the 4 transcripts, indicating that it is generally sufficient to cluster slow codons at the beginning of the transcript rather than ordering them explicitly in the form of a ramp. The one exception to this is transcript 4, whose ramp form has much higher J/θ than its throttled form; this is due to the severely rate limiting codon at the end of the throttle region.

Importantly, these 4 transcripts also exhibit a wide range of "tunability" by changing initiation rates α. Unlike the "flat" and "step" transcripts, transcripts 1, 2, and 4 do not exhibit increased J/θ when α is decreased. This means that for these transcripts, a high ribosomal efficiency *must* be achieved by optimizing codon order. This observation is intuitively understandable when one considers the presence of "bot-

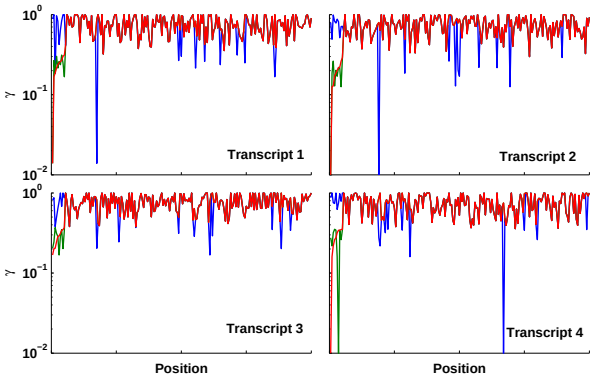
tleneck" codons on these transcripts–because the effect of a bottleneck is to decrease the critical value α_c at which the transcript reaches saturation-occupancy, this means that θ is insensitive to changes in α on a much wider range of parameter space.

Figure 3: The effect of re-ordering codon efficiencies.

(a) Translational optimality of codon-reshuffled mRNAs. Plotted are simulation results for 4 randomly generated transcripts of $N = 200$ and their optimized variants, at different initiation rates in the range $0.01 \leq \alpha \leq 30$. Initiation rates aren't indicated explicitly, but increasing initiation rate roughly increases J and decreases J/θ on the plot. Transcripts 1, 2, and 4 have much lower J than transcript 3, an observation explained by the presence of "bottleneck" codons (Fig. 3b).



(b) Codon efficiency profiles for 4 randomly generated mRNA transcripts. The plots represent $\log \gamma_i$ versus i. All but one of the transcripts contain a "bottleneck" slow codon. See Fig. 3a for color key.



Conclusions

Codon efficiency "ramps" have recently been proposed as a mechanism by which cells optimize translational efficiency. However, as I show in my simulations, improving the efficiency of translation with respect to ribosomes requires a decrease in the overall flux of proteins off a transcript. This is true even when the ramp is not formed by decreasing codon efficiencies on an otherwise maximally fast transcript, but even when the ramp is built by reordering jump rates on an existing (suboptimal) mRNA. The latter can be seen as a plausible mechanism by which ramps arise from mutation–for example if an amino acid appears both at the beginning and middle of a gene, the middle codon can mutate to a more efficient version while the beginning codon mutates to a less efficient one.

Furthermore, even if it were advantageous to trade mRNA efficiency for ribosomal efficiency, this can be achieved by adding a "throttle" region as well as by decreasing the initiation rate of ribosomes. Given that codon ramps exist in actual cells, this implies that there are constraints on the evolutionary plasticity of the initiation mechanism, or that these transcripts are "forced" to contain bottleneck codons for an unrelated biological reason. Even then, however, increased ribosomal efficiency doesn't come without some cost to mRNA efficiency, indicating that the distribution of costs between different facets of gene expression is a highly nontrivial task for cells.

References

Akashi, H. (1994). Synonymous Codon Usage in Drosophila melanogaster: Natural Selection and Translational Accuracy. *Genetics*, 136:927–935.

Derrida, B., Domany, E., and Mukamel, D. (1992). An exact solution of a one-dimensional asymmetric exclusion model with open boundaries. *Journal of Statistical Physics*, 69(3-4):667–687.

Drummond, D. A. and Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2):341–52.

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 E. coli translational system. *Journal of molecular biology*, 151(3):389–409.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., and Weissman, J. S. (2009). Genome-wide analysis

in vivo of translation with nucleotide resolution using ribosome profiling. *Science (New York, N.Y.)*, 324(5924):218–23.

Janowsky, S. a. and Lebowitz, J. L. (1994). Exact results for the asymmetric simple exclusion process with a blockage. *Journal of Statistical Physics*, 77(1-2):35–51.

Kolwankar, K. and Punnoose, A. (2000). Disordered totally asymmetric simple exclusion process: Exact results. *Physical Review E*, 61(3):2453–2456.

Krug, J. (2000). Phase separation in disordered exclusion models. *Brazilian Journal of Physics*, 30(1):97–104.

Lakatos, G. and Chou, T. (2003). Totally asymmetric exclusion processes with particles of arbitrary size. *Journal of Physics A: Mathematical and General*, 36(8):2027–2041.

Navon, S. and Pilpel, Y. (2011). The role of codon selection in regulation of translation efficiency deduced from synthetic libraries. *Genome biology*, 12(2):R12.

Schütz, G. and Domany, E. (1993). Phase transitions in an exactly soluble one-dimensional exclusion process. *Journal of Statistical Physics*, 72(1-2):277–296.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141(2):344–54.

Uemura, S., Aitken, C. E., Korlach, J., Flusberg, B. a., Turner, S. W., and Puglisi, J. D. (2010). Real-time tRNA transit on single translating ribosomes at codon resolution. *Nature*, 464(7291):1012–7.