Using the Totally Asymmetric Exclusion Process as a Model for Protein Translation

USING THE TOTALLY ASYMMETRIC EXCLUSION PROCESS AS A MODEL FOR PROTEIN TRANSLATION

BY

PAK LAM (PHILIP) LEE, B.Sc. (Honours)

A THESIS

SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2012)	McMaster University
(Physics and Astronomy)	Hamilton, Ontario, Canada

TITLE:	Using the Totally Asymmetric Exclusion Process as a Model for Protein Translation
AUTHOR:	Pak Lam (Philip) Lee B.Sc., (Mathematics and Physics) University of Toronto, Toronto, Canada
SUPERVISOR:	Dr. Paul G. Higgs
NUMBER OF PAGES:	xvi, 94

dedicated to my family, aunts, uncles, cousins, and grandparents.

Abstract

This thesis details the development of a kinetic model of translation which takes into account codon usage. The process of translation involves ribosomes decoding a sequence of codons to produce a protein. Codon usage is important in the kinetics of translation since experiments have shown that codons are processed at different rates. Codons which code for the same amino acid appear with unequal frequencies and certain synonymous codons are preferred by high expression genes. The relationship between translational efficiency and codon adaptation is explored in this thesis.

We use a simple physics model called the totally asymmetric exclusion process (TASEP) to emulate the action of ribosomes, and the decoding of mRNA in protein elongation. The simple model is parameterized by an initiation rate that determines how quickly new ribosomes are introduced onto the lattice, and the rate of motion for ribosomes associated with a site on the lattice (codon message). Based on bioinformatics studies, we assign codon speeds so that codons preferred by high expression genes are translated more quickly.

The model captures important aspects of translation like ribosome collision and codons of different speeds, and simulating it allows us to see details in dynamics which are inaccessible to experiments. TASEP has non-trivial behaviour when codon rates, and the rate of ribosome binding is varied. Slow codons can cause ribosomes to pause and may lead to a queue. We approximated real genes with its average rate, and with its slowest codons to test the salient features of how codons are used on mRNAs. We found that codon selection is important in determining when queues occur, and the ribosome density on genes. The model also shows that highly expressed genes queue later than low expression genes. The simple model gives us general insights into the translational selection of codons, and the important kinetic parameters.

Acknowledgements

I would like to give thanks to my supervisor Dr.Paul Higgs for allowing me to work on this project, and his excellent academic guidance. His patience and support throughout the project has made it possible for me to complete this thesis. I truly admire and benefitted from his positive attitude and integrity in scientific research. Working in Dr.Higgs' group has taught me new and wonderous things in genetics and biology.

I would also like to thank the professors and students in the physics and astronomy, and material science faculties at McMaster, and all the scientists I have encountered in my studies. Their dedication to science, and research experience has taught me valuable lessons about life in science, and I am glad to have learned from them. Thanks to the administrative, and computing staff who have been so dilligent and helpful with applications and resources. I also would like to acknowledge the university's important support in scientific endeavours.

All that I have is from my family, without whom I am nothing. Their unconditional love and support has made me a wiser person. This thesis is dedicated to them for their faith in me.

Contents

List of Figures

Chapter 1

Introduction to Translation and Codon Usage

1.1 Ribosomes and translation

The ribosome is ubiquitous in cells, and its function is conserved over all organisms. The ribosome is a quaternary structure of folded RNAs and proteins, involved in the transfer and linking of amino acids into polypeptide chains. It uses a sequence of genetic information in the form of codons on an mRNA to make a a protein that performs the various functions in a cell. Each codon encodes a specific amino acid in the polypeptide chain. The mRNA sequence is translated by the ribosome into a protein one codon at a time with high accuracy. It is observed that some codons are used more often than others, and that natural selection would prefer more efficiently translated codons. There is also experimental evidence that codons are translated at different rates (Curran and Yarus, 1989; Sørensen and Pedersen, 1991; Sørensen et al, 1989). Therefore, the pattern of codon usage will influence the kinetics of translation. The kinetic picture of translation to be developed here will give a higher level of detail of how codons are adapted in a cell to optimize translation. Codon usage varies among genes, and can also vary within the same gene. The variation of codon usage within one gene can give rise to features such as clusters (blocks of rare codons), and a ramp (rare codons used at the beginning of the gene). There is a bias in the codon composition at the beginning of mRNAs (Bulmer, 1988; Tuller et al, 2010). The configuration of codon speeds on an mRNA produces non-trivial motion of ribosomes on an mRNA, so that we are required to model their motion as proteins are synthesized during translation. The cell is producing proteins at all times in a cell, and hence the number of ribosomes on an mRNA is not conserved, and there is always a flux of translating ribosomes. The configuration of codons on an mRNA causes ribosomes to proceed at different speeds, and when multiple ribosomes are engaged in translation, could lead to queueing. When there are a sufficient number of ribosomes on an mRNA, the rate of protein production will have non-trivial dependence on the local, codon determined elongation rate. A physics model which describes a queueing process in a highly dynamical and out-of-equilibrium system such as in a cell is used in this thesis to model translation. The main purpose of this thesis is to combine bioinformatics studies of codon usage with a dynamical model in physics to describe the production of proteins in a cell.

1.2 Details of ribosome function

Following is a summary of a review by V. Ramakrishnan (2002). The ribosome is a macromolecular complex composed of two subunits, measured in svedberg sedimentation units, in prokaryote the 30S (40S eukaryote), and the 50S (60S eukaryote) subunits. The ribosome itself has 70S at a diameter of $\sim\ 200\,\mathring{A}$ and 80S at $\sim\ 250 - 300\,\mathring{A}$ for prokaryote and eukaryote respectively. The subunits are quaternary structures made of serveral long ribosomal RNAs and many ribosomal proteins. The two subunits bind onto an mRNA to initiate translation.

Initiation onto an messenger RNA (mRNA) chain begins at the 'Shine-Dalgarno' sequence in prokaryotes, or the 'Kozak' sequence in eukaryotes. The subunits recognizes these codon regions, and the ribosome is assembled there. The first codon translated by a ribosome is the start codon, which encodes the amino acid formylmethionine (fMet). fMet is a special amino acid that is the leading peptide of proteins. The start codon is the same one that codes for methionine, but its position at the beginning of the gene allows it to only select the starting peptide fMet. By chemical means, we infer that there it is aided by intiation factors (IF1, IF2, and IF3) that cause conformational changes to the ribosome to select the fMet starting peptide tRNA, and prevent binding of other tRNAs during this phase. tRNAs carry the amino acid to be added to the protein chain.

A ternary complex of aminoacyl transfer RNAs (aa-tRNAs, including Met-tRNA), elongation factor Tu (EF-Tu), GTP is brought into the ribosome's A (accomodation) site. The tRNA structure contains an anti-codon which pairs with a codon on an mRNA. In the presence of an aa-tRNA, the ribosome decodes the mRNA by rejecting those tRNAs that do not form the proper codon-anti-codon pair with the codon at the A site. When a correct tRNA is bound to the lower subunit A site, GTP is hydrolyzed away by EF-Tu, and the EF-Tu dissociates from the tRNA ternary complex. During this action a peptidyl transfer occurs. The growing peptide is transfered from the tRNA bound at the next site, the peptidyl (P) site to the tRNA at the A site. Elongation factor G (EF-G) translocates the ribosome and the pair of mRNA bound tRNAs will be shifted from the A site to the P site, and P site to the E (exit) site in the ribosome. Spent aa-tRNAs will exit the ribosome. The tRNAs will now have moved relative to the ribosome. The numerous reactions and steps involved in peptide elongation is depicted in figure 1.1.

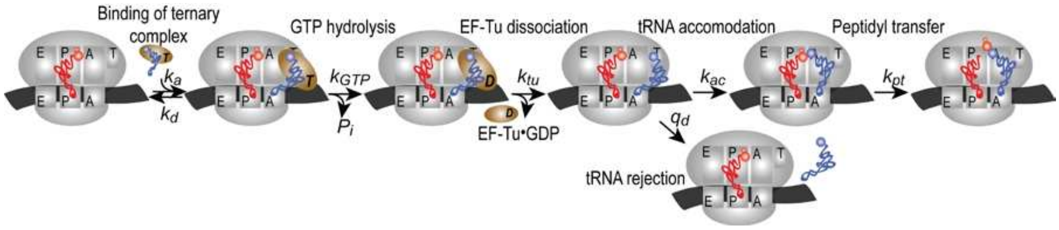


Figure 1.1: There are conjectured and experimentally verified a large number of steps to the elongation of a peptide. Source:*Ehrenberg M. (2009) Scientific Background on the Nobel Prize in Chemistry 2009 - Structure and Function of the Ribosome. The Royal Swedish Academy of Sciences.*

The ribosome terminates translation when the protein is complete and it reads a stop codon on the mRNA. The stop signal encodes release factors that induce hydrolysis of the peptidyl link at the P site tRNA, allowing the polypeptide to be release from the ribosome. This allows the completed protein to fold freely into its functional form. A recycling release factor and EF-G disassemble and reset the ribosome for future use.

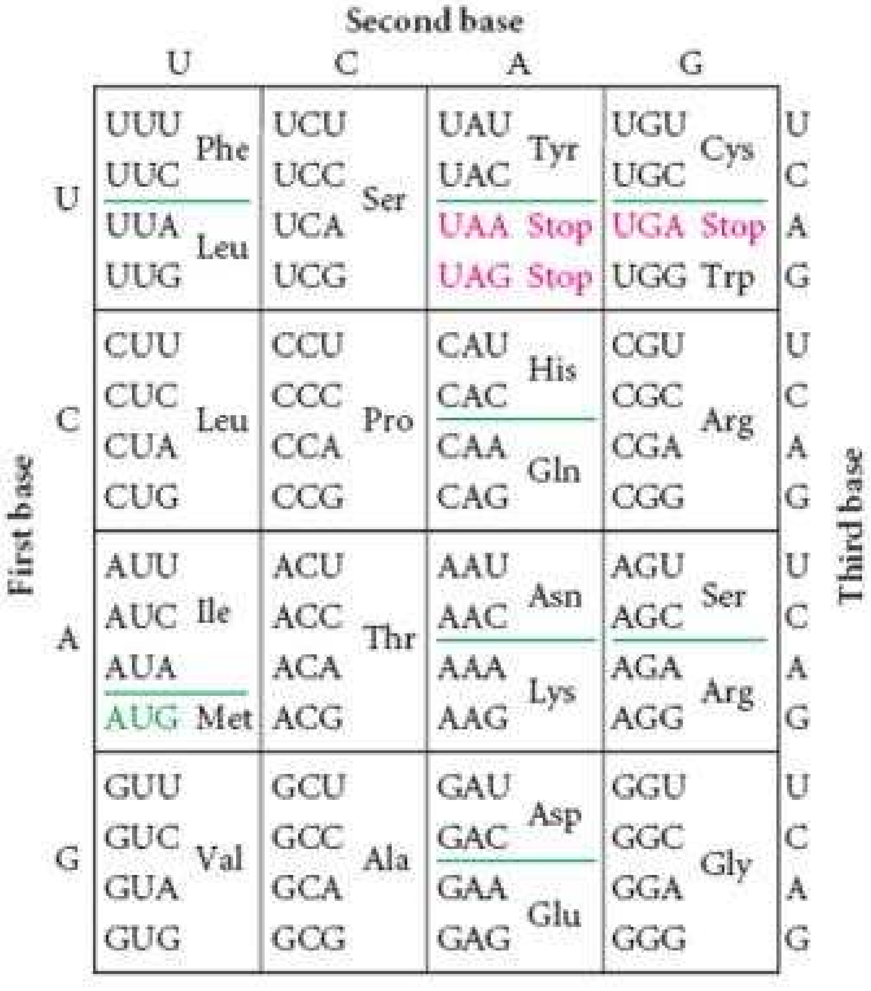
In this thesis, we will use the Totally Asymmetric Simple Exclusion Process (TASEP) model to describe the dynamics of the ribosome. This model has three rates, corresponding to initiation, elongation, and termination. In the model each of these is treated as a single step, and many of the above details are omitted. The initiation will happen at a rate (α) representative of a ribosomal binding and assembly event. So that a ribosome binds and assemble on to an mRNA in only one step. While polypeptides are not simulated, their elongation corresponds to a ribosome moving one codon further along an mRNA at a rate denoted by v_x. The rate at which elongation happens is determined by the position of the ribosome on the mRNA, or which codon it is processing. This rate should be assigned based on genetic and cellular parameters such as tRNA concentration, and codon usage (Kurland, 1991; Akashi, 2003). The termination in our picture is unimportant, and will always occur when a ribosome hops to the end of an mRNA.

1.3 Bioinformatics evidence for translational selection

Synonymous codons are translated into the same amino acid. Different codons can translate into the same amino acid because the genetic code is degenerate. Specifically, there are 64 possible codons, and only 20 amino acids. The codons and their corresponding amino acid are listed in the table in figure 1.2. It is observed that codon frequencies for synonymous codons are not equal (Sharp and Li, 1987). While mutation may cause this bias, it is seen that codon frequencies in high expression genes differs from that in low expression genes. This suggest that selection is present in high expression genes (Higgs and Ran, 2008; Sharp et al, 2005). The fitness of an organism depends on the efficiency of translation as translation is an important step in the growth of organisms. During the growth of organisms, such as bacteria, an individual mRNA needs to be translated multiple times before division. It has been observed that strong codon usage bias in an organism's mRNA correlates with the growth rate of an organism (Akashi, 2003; Dos Reis et al, 2003; Sharp et al, 2005). So that the usage of codons is a mechanism through which selection works on translation. Further more, tRNAs play an important role in translation: they bring in amino acids that form proteins. Higher tRNA concentration generally leads to quicker translation of the coded amino acid. It has been observed that tRNA abundance and codon frequencies have co-evolved to improve translational efficiency (Ikemura, 1981; Percudani et al, 1997; Duret, 2000). Since these processes happen constantly in a cell, any sub-optimization in translational efficiency accumulates and decreases the fitness of an organism. On the other hand, efficient translation is compounded, and would be selected for in general.

1.4 Experimental work done on translation

The difficulty in studying the ribosome *in vivo* is due to limited spatial and temporal resolution of the large number of steps, conformational changes, reactions, and reactants known to be involved in translation. Figure 1.1 depicts possible steps in the elongation of a peptide. Despite the difficulty in this endeavour there are a few



	Second base								
First base	U		C		A		G		Third base
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Figure 1.2: This table shows 64 codons and their associated amino acid. Source:*Smith A. (2008) Nucleic acids to amino acids: DNA specifies protein. Nature Education 1(1)*

results which can guide the development of our theoretical model for translation.

1.4.1 Codon rates experiment

Our model for translational kinetics requires as a parameter the elongation rate for codons. It is important to recognize that in a cell, codons are translated with different speeds. By measuring the incorporation of labelled methionine into completed proteins at different times, the following experiments estimated the rates different parts of an mRNA is processed by a ribosome. It was shown by Sørensen et al (1989) that genes inserted with infrequent codons are translated more slowly than genes with frequent codons. In that work, they concluded that frequently used codons and rare codons may have a six-fold difference in their rate of translation. More precise values for the translation rate of codons were obtained later by Sørensen and Pedersen (1991). They used specific gene inserts that allowed them to compare the translation rates of two glutamic acid codons GAA, and GAG in *Escherichia coli.* These two codons were chosen because they are synonymous codons (translating the same amino acid), and that they are recognized by the same tRNA. Any difference in translation speed in this case cannot be attributed to difference in tRNA abundance, and that it is the difference in the way synonymous codons are recognized. They found that GAA was translated with a rate of 21.6 codons per second, which the codon GAG was translated about 3.4 times more slowly at a rate of 6.4 codons per second. In addition to measuring translation rates for codons translated by the same tRNA, the experiment also found indirect evidence for queueing. When an insert to the gene cause ribosome motion to be so slow on the mRNA, that queueing causes codons further along to be infrequently translated by ribosomes.

1.4.2 Ribosome density mapping

An mRNA with one or more initiated or translating ribosome is called a polysome. Electron microscopic images of polysomes are show in figure 1.3. As seen in the electron microscope image, there appears to be multiple ribosomes translating an individual mRNA. The linear arrangement suggested to the authors that this is indeed a polysome. The polysomes were collected from *E. coli* K-12 bacteria that were induced to produce β-galactosidase, an enzyme in the cell. A closer look at polysomes was done by Arava et al (2005) using ribosome density mapping. Ribosome density mapping (RDM) is a technique which uses cleft polysomes to give quantitative information about the density profile of ribosomes along an mRNA during translation. This technique provides *in vivo* insights into translation and its controls (Arava et al, 2005). A schematic for the experiment is presented in figure 1.4. For a small number of bound ribosomes, the experiment is able to resolve at ±1 ribosome.

Arava et al (2005) mapped the ribosome density of GCN4 mRNA in yeast to explore three main controls of translation: initiation, elongation, and termination. An mRNA was cleft roughly in half with the observation that each half having equal density. This is a necessary indicator for an mRNA with uniform density. Cleaving a third off both ends gives stronger indications to other translational controls. For the mRNA investigated, the ends had similar ribosome densities. This implies that ribosomes tend not to dissociate and fall off the mRNA during elongation, that elongation processivity is generally high (at least for the mRNA investigated). From this,

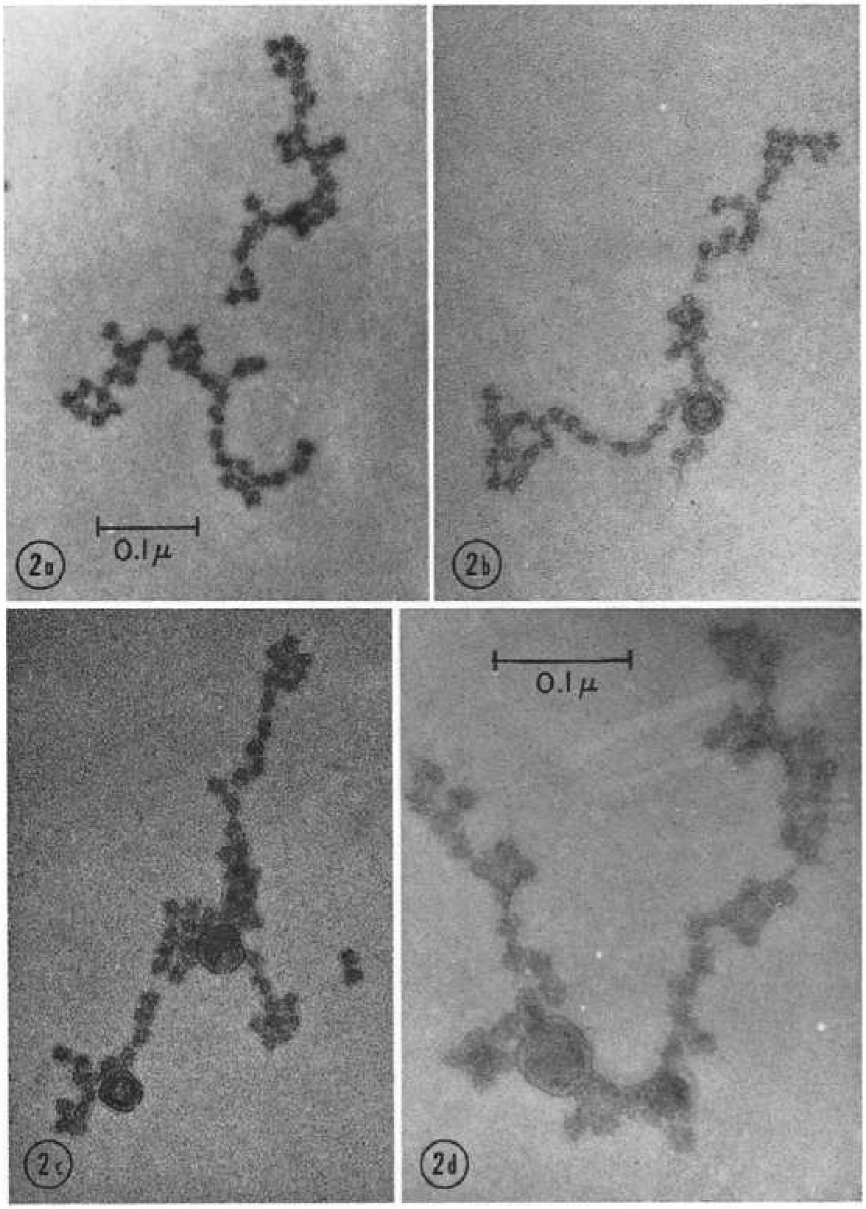


Figure 1.3: Ribosomes on mRNA. The chain of black spots are ribosomes. Source: *Slayter H, Kiho Y, Hall CE, Rich A. (1968) An electron microscopic study of large bacterial polyribosomes. J Cell Biol., 37, 583590.*

we will model translation without ribosome dissociation.

Because a pile-up of ribosomes at the end of the mRNA is not detected, it means that the termination rate do not limit processivity. As long as ribosomes remain associated on the mRNA during translation, a current of ribosomes coming from the open reading frame (ORF) is maintained, which is necessary for a pile-up to occur. Since if dissociations were common then it would weaken the current of ribosomes from the beginning of the mRNA. This has informed our model in that the termination rate is always set to be high relative to initiation and elongation rates.

1.4.3 mRNA structure

Other factors that can affect translational efficiency include mRNA structure. An effect where mRNA length and ribosome density are inversely correlated was observed by Arava et al (2003). A possible explanation was that longer mRNAs are less efficient initiating ribosomes, because longer strands are more prone to mRNA secondary structures that can inhibit initiation. The stability of mRNA secondary structure at ribosome binding sites was found to affect the level of gene expression *in vivo* (de Smit and van Duin. 1990). In Gu et al (2010), the relationship between structure formation and codons usage at the beginning of mRNAs is explored, with the general result that RNA structure formed by nucleotide interactions there tend to be less stable. In any case, it appears that shorter mRNAs having a higher ribosome density is due to translational regulatory mechanisms. While mRNA length is not explored in our model, the use of specific codons at the beginning of mRNAs affects the codon dependent translation rate there.

1.5 Measuring codon adaptation

As seen from experiments summarized in section 1.4.1, a ribosome processes codons at different rates. The usage and order of codons on an mRNA determines the end product protein, and the rate at which the proteins are produced. The codons' adaptation is hypothesized to depend on cellular parameters such as the relationship between codon frequency and expression in the genome, and the abundance of tRNAs that recognizes the codon. Following are some proposed methods to measuring codon adaptation for a specific set of translation apparatus.

1.5.1 Codon adaptation index, CAI

Synonymous codons are used with unequal frequency, that is, even though codons are translated into the same amino acid, some appear more frequently than others. The codon adaptation index developed by Sharp and Li (1987), can be used to assess codon adaptation in relation to a set of highly expressed genes. It is used to compare the usage of synonymous codons between different genes in a genome. The codon adaptation index is defined for a gene of length L,

$$CAI = (\prod_k w_{i_k})^{1/L}$$

where i_k denotes the codon at position k along the mRNA, and w_{i_k} are the relative adaptiveness of a codon based on a set of highly expressed genes. The relative adapativeness of a codon is calculated based on a set of highly expressed genes and is

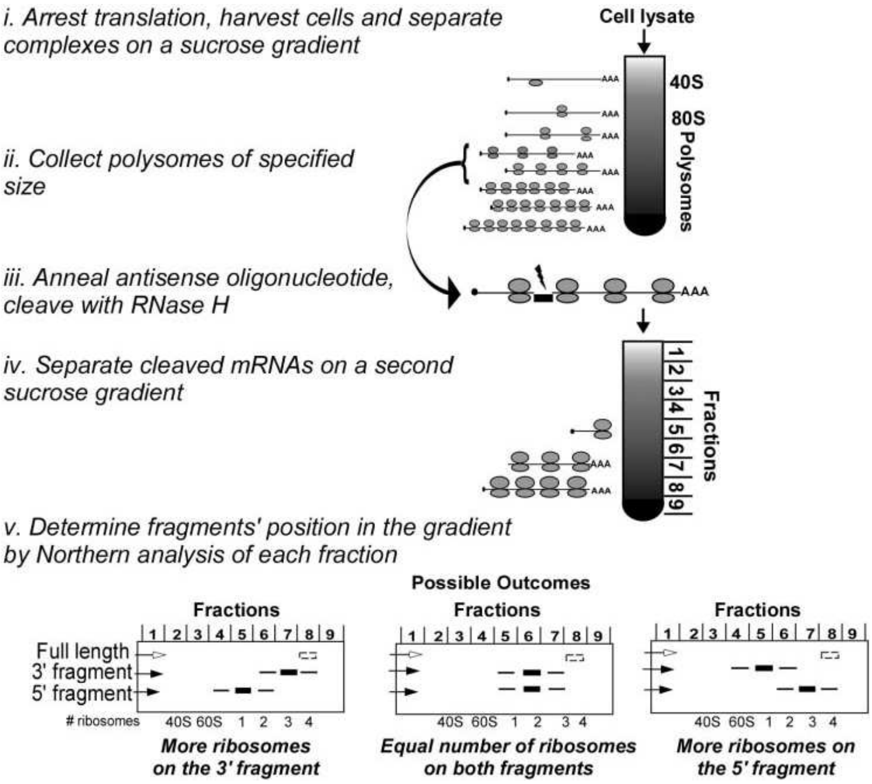

i. Arrest translation, harvest cells and separate complexes on a sucrose gradient
Cell lysate
40S
80S
Polysomes
AAA
ii. Collect polysomes of specified size
iii. Anneal antisense oligonucleotide, cleave with RNase H
iv. Separate cleaved mRNAs on a second sucrose gradient
Fractions
1 2 3 4 5 6 7 8 9
v. Determine fragments' position in the gradient by Northern analysis of each fraction
Possible Outcomes
Fractions
Full length
3' fragment
5' fragment
# ribosomes
40S 60S 1 2 3 4
More ribosomes on the 3' fragment
Equal number of ribosomes on both fragments
More ribosomes on the 5' fragment

Figure 1.4: Schematic for ribosome density mapping from [RMD]

defined as,

$$w_i = \frac{x_i}{x_{max}}$$

with x_{max} being the number of the most used codon for the amino acid in the reference set, and x_i are the occurrences of the codon i for that same amino acid in the reference set. In this case, the most used codon in a synonymous set has a relative adaptiveness of 1, and those less frequent have values $w_i < 1$. The set of reference genes are usually chosen from those of ribosomal proteins and elongation factors, since these are the main components of the translation apparatus, and should appear most often in the cell. The w_i values in *E. coli* are given in table A.3 (as ϕ_H) and will be discussed in chapter 3. A gene with a high CAI frequently uses those codons preferred in the reference set. This geometric average can predict the expression level of a gene, as well as compare codon usage between different organisms. CAI of genes show bias in the usage of synonymous codons between high and low expression genes; this bias has been studied using a evolution selection model (Bulmer, 1991). A kinetic model of mRNA translation can give a closer look at the way codons are used in a sequence.

1.5.2 tRNA adaptation index, tAI

tAI is a measure of translational selection for genes based on tRNAs' gene copy number for a given codon (dos Reis et al, 2004). The ribosome facilitates this recognition, and its kinetics is determined by the strength of this pairing along with other biochemical signals (Ninio, 2006). The more abundant a tRNA, the more likely the codons it pairs with can be recognized by a ribosome, and the more quickly its corresponding codon can be translated. This measure of translation efficiency takes into account Watson-Crick pairing as well as interaction at the wobble position of a codon.

The strengths of these pairings is a chemical quantity, and it is observed that tRNA-codon pairings that are Watson-Crick pairs are more frequent with preferred codons (Ikemura, 1981). The tAI is defined as the geometric average of relative adaptiveness for the codons in a gene. The absolute adaptiveness for a codon i is defined as,

$$W_i = \sum_j b_{ij} Nj$$

summing over tRNAs which recognizes the ith codon. b_{ij} are selective constraints based on the efficiency of codon i and tRNA j pairing, and N_j are gene copy numbers for the jth tRNA. Typically, a tRNA's gene number correlates with the tRNA's abundance in a cell. The relative adaptiveness w_i for codon i is just the absolute adaptiveness W_i normalized by absolute adaptiveness of the best adapted codon W_{max} in the tRNA pool,

$$w_i = \frac{W_i}{W_{max}} \tag{1.1}$$

when the absolute adaptivness is non-zero, and w_{mean} if the absolute adaptiveness vanish. The tAI of a gene is then,

$$tAI = (\prod_k w_{i_k})^{1/L}$$

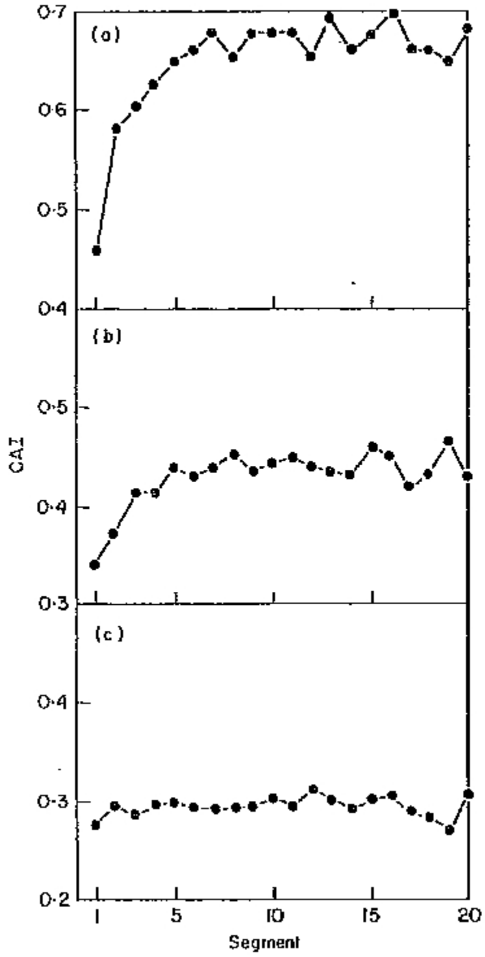
where the product k is over the length of the gene L long. Relative tRNA abundance is a quantity that is selected in evolution. It is important in determining the overall kinetics of ribosomes in translation. Tuller et al (2010) used a variant of the tAI to investigate the translation problem. They assumed that tRNA concentration as the most salient determinant in ribosome kinetics. Table A.3 shows the tAI values for different codons, and these will be discussed in chapter 3.

1.6 Biological motivations

The main biological theme in this thesis is to investigate how codon usage may affect translation. The density profile of ribosomes on mRNAs depends on parameters which are kinetic in nature. These are the rate of ribosome initiation onto an mRNA, and the elongation rate that depends on the codon being translated. One of the main questions we will ask is whether the presence slow codons in mRNAs causes queueing of ribosomes. We are trying to distinguish the roles of initiation or the kinetic binding of ribosomes, and elongation or codon speed in the production of proteins. In the framework of the following model, there are two cases to consider: initiation limited, and elongation limited.

For the case where codon speed limits protein translation, there can be two types: a queue of ribosomes, or a blocking of ribosomes from binding. In Sørensen and Pedersen (1991), there was evidence that queueing occurs when a gene is modified to have a sufficiently long and slow stretch of codons. Since it is hard to obtain experimental data at the time and length scale for translation, and the variety of codon configurations, we would like to investigate with this model quantitatively the effects of a slow codon cluster on translation. Rare codons are found at the beginning of genes (Bulmer, 1988; Tuller et al, 2010). Findings from Bulmer (1988) is shown in figure 1.5. However, this data should not suggest that this is an expression regulatory mechanism since it occurs at all expression levels. Slower codons at the beginning of mRNAs increases the time any ribosome spends there, and a ribosome will block other ribosomes from binding onto the mRNA. Tuller et al (2010) discusses how this can prevent ribosomes from piling up and queueing further along the mRNA. This may be favourable in a cell with finite resources since the time that ribosomes are sequestered idle on the mRNA in queues is decreased and would increase ribosome availability in the cytoplasm (Kurland, 1991). The phenomena we wish to explore is the effect of codon usage on translation, a process which is heterogeneous (works in many different genes), and the interactions (ribosome collision, queueing) are hard to detect experimentally.

Initiation is the other mechanism by which translation can be regulated. When the initiation rate is low, the time between ribosomes may be longer than any pause at slow codons. This means that no significant queues are possible when the initiation rate is low since it means that ribosome density is also low. The 'ramp' mechanism mentioned in (Tuller et al, 2010) is also ineffective when initiation rate is low, because there are not enough ribosomes to allow queueing in the first place. As discussed in (Bulmer, 1991), the elongation rate or the average choice of codons on mRNAs is important only when initiation is not the rate-limiting step in translation. While the usage of fast codons may generally speed up translation, it only comes to affect translation when the initiation rate is high. Others have suggested that this signal of codon bias at the beginning of genes is that due to secondary structures (Gu et al, 2010). Experiments have indicated that mRNA structure near the ribosomal binding site play a predominant role in expression (Kudla et al. 2009), and less stable structures there lead to more efficient ribosome initiation (de Smit and van Duin, 1990). This could mean that codons are selected to minimize structure at the beginning of mRNAs, and that initiation is an important parameter in protein production and gene expression. By varying the initiation rate in our model, we would like to look in



[H]

Figure 1.5: This is a plot of the gene set average of CAI (codon adaptation index) for segments of 15 codons, versus the position on a gene. The value of CAI (codon adaptation index), is lower at the beginning of genes which have (a)high, (b)medium, and (c)low expression. While high expression genes generally use common codons, the beginning of genes on average uses rare codons. This use of slow codons at the beginning is present for all types of genes. Source:*Bulmer, M. (1988). Codon Usage and Intragenic Position. J. theor. Biol.. 133, 67-71*

more detail its effect on the dynamics of translation.

The relationship between codon usage, mRNA configuration, and ribosome initiation is not a simple one. These are interfering processes which have complex interactions with each other, and the model developed in this thesis intends to clarify some of these issues.

1.7 Definition of the TASEP model

The Totally Asymmetric Simple Exclusion Process (TASEP) is a non-equilibrium statistical mechanics model that describes particles hopping along a lattice. It hase been used as a model of driven lattice gas (A. B. Kolomeisky et al, 1998), traffic jams (K. Nagel and M. Schreckenberg, 1992), for transport processes in biophysics including ribosome dynamics during translation (C. MacDonald et al, 1968; C. MacDonald and J. Gibbs, 1969), the dynamics of RNA polymerases during transcription (Tripathi et al, 2009), and the motion of motor proteins along microtubules (A.B. Kolomeisky, 2001).

An important variation of TASEP is the open uniform TASEP with particles of size 1, this was solved by Derrida et al (1992). In this model, particles move along a one-dimensional lattice of L sites. A particle may enter site 1 at rate α if this site is not occupied. A particle at site i may move to the next site at rate 1 if the next site is not occupied. A particle on site L may leave the lattice at rate β. We call this the open uniform TASEP because the number of particles is not constant during its dynamical evolution as particles are allowed to enter and leave. The fundamental quantities that are of interest to calculate in the TASEP model are the current, J, which is the number of particles exiting the lattice per unit time, and the density, ρ, which in the case where particles are of size 1, is the fraction of occupied sites. The results of Derrida et al, 1992 will be summarized in section 2.2 below.

In our TASEP model for translation, each site of the one-dimensional lattice is a codon on messenger RNA, and each particle is a ribosome which hops in only one direction along the lattice. The footprint of a ribosome on an mRNA is larger than one codon. Thus we will consider ribosomes as extended objects of length l. There have been various estimates of l. In an experiment by M. Takanami and G. Zubay (1964) radioactive poly U templates were acted on by ribosomes and then digested. They gave an estimate for the ribosomal footprint in *E. coli* to be about 27 residues, or 9 codons on an mRNA. Mitarai et al, 2008 ran a kinetic model of translation using a ribosome size of 11 codons, or 33 nucleotides based on a study of ribosomal initiator sequence (J.A. Steitz, 1969). Ribosomes are excluded from one another by not allowing any pair of ribosomes to be closer than l lattice sites. So that if a ribosome is at site i, the range $i+1$ to $i+l-1$ cannot contain any other ribosomes. A ribosome may move provided that the ribosome in front of it is more than l sites along the lattice, or further along than position $i+l$, otherwise it must wait for the ribosome in front to move before continuing. A new ribosome may be initiated at rate α onto the lattice when there are no ribosomes in the first l sites, and a ribosome is terminated at rate β when it is at the end of the lattice. A current J of ribosomes is defined for ribosomes that has completely moved through the lattice per unit time. If there are n ribosomes on an mRNA of length L, we will define the coverage density as $\rho = \frac{nl}{L}$. These are emergent properties in TASEP found in its dynamic steady state. The TASEP with extended objects has been studied previously

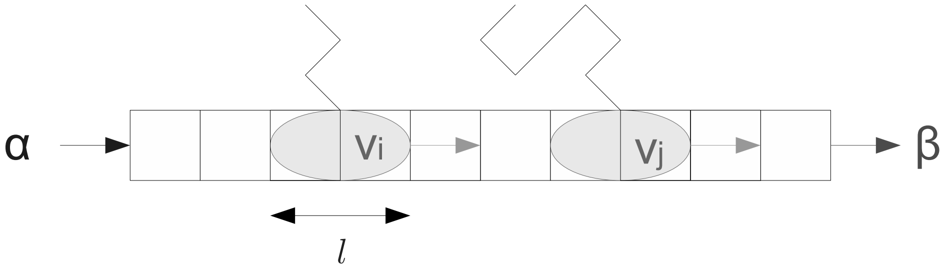


Figure 1.6: The totally asymmetric exclusion process (TASEP). Objects of length l are injected at a rate of α, hops with rates of v_i's, and exits at a rate of β. Also depicted are proteins elongated by the ribosomes.

by (Lakatos and Chou, 2003) and (Shaw, Zia, and Lee, 2003), and we will summarize the important results for this case in section 2.3. A schematic of the process is shown in figure 1.6.

In the biological context, as the ribosome moves, a protein is elongated, and a current of proteins is produced. This is our model of protein production involving the kinetics of ribosomes.

The other important feature that needs to be added in order to use the TASEP for translation is that the ribosome deals with different codons at different rates, depending on the availability of the corresponding tRNAs and the nature of the codon-anticodon interaction, as discussed in section 1.5. We will therefore consider non-uniform TASEPs in which each site has a hopping rate v_i which depends on the codon at position i. The cycle of elongation is quite complex as discussed in sections 1.2 and 1.4. In our simulations, we will make the simplification of treating each elongation cycle as a single step, but the rate of this step may be different at each site, depending on what the codon is at the site. We will discuss previous work on non-unniform TASEPs in sections and present our own simulations of simple non-uniform cases in chapter 2. When we wish to do realistic simulations of real gene sequences, it is necessary to assign a hopping rate to each codon. The way we chose to assign these rates for real sequences will be discussed in section 1.7.

1.8 Aims of this thesis

The work presented in this thesis attempts to use simulations to model protein translation kinetics. In chapter 2, we present and test several theoretical predictions for TASEP models with a cluster of slow codons in the sequence, and a ramp of slow codons at the beginning of a sequence. A cluster of slow codons is where there is a sequence of codons that requires a longer time for the ribosome to pass through than the average. This causes ribosomes to pause and is used to describe ribosome interactions in the form of queueing during translation. A ‘ramp’ is a sequence of slow codons at the beginning of the system, and this causes ribosomes to spend more time at the beginning of sequences, increasing ribosome interaction with the initiation process. Theory on the non-trivial behaviours of TASEP are presented, and we used a stochastic simulation to test the theoretical predictions. In addition to presenting a theory on TASEP with a single inhomogeneity, we extended TASEP theory with a single inhomogeneity to a more useful form that approximates for up to l slow codons. These concepts are applied in chapter three for mRNA translation and translational selection.

The kinetic model developed in this thesis allows us to test how bioinformatics studies of codon frequencies can determine translation kinetics. We attempt to find a way of setting codon hopping rates that is consistent with bioinformatics studies of codon frequencies. We simulated real gene sequences from *E. coli* in chapter three based on a bioinformatics measure of codon selection. Since real genes are highly heterogenous in composition, we develop theoretical approximations to investigate features of codon distribution that are important to the translation of *E. coli* genes. We were able to quantitatively determine how well the theoretical approximations can predict the translation behaviour of real gene sequences for this model.

Chapter 2

Testing the TASEP model for simple artificial codon frequencies

2.1 Implementation of the TASEP model using the Gillespie algorithm

The Gillespie algorithm is a standard method of simulating the dynamics of stochastic processes such as chemical reaction systems (D.T. Gillespie, 1977). Here we will use the Gillespie algorithm to implement the TASEP model.

We number the ribosomes currently on the mRNA as $k = 1, 2, \ldots, n$, where n labels the ribosome that has advanced the furthest and 1 labels the ribosome most recently added on to the mRNA. Let $i(k)$ be the site occupied by ribosome k. Let r_k be the hopping rate of ribosome k. If the ribosome is not blocked, $r_k = v_{i(k)}$, and if it is blocked by the preceding ribosome, $r_k = 0$. If a ribosome reaches the last site, its rate of motion is β. The rate of adding an additional ribosome ($k = n + 1$) at site 1 is $r_{n+1} = \alpha$, if there is no ribosome on the first l sites, otherwise $r_{n+1} = 0$. The sum of the rates of all possible events is,

$$R = \sum_{k=1}^{n+1} r_k.$$

According to the Gillespie algorithm, one random event, k, is chosen with probability proportional to its rate: $p_k = r_k/R$. By randomly generating a number between 0 and R, an event, in this case, either ribosome initiation, hopping or termination is picked. Once the system is updated, the dynamics has evolved by a time τ chosen from an exponential distribution with a mean $\frac{1}{R}, i.e.$

$$P(\tau) = Re^{-R\tau}.$$

If a simulation begins with no ribosomes bound to the mRNA, it takes some time for the system to reach dynamic steady state. In our simulations, we waited until 1000 ribosomes had completely translated the mRNA before measuring the steady state quantities of interest.

In order to measure the current J, we simply keep track of the total number of ribosomes that exited the system, and the total dynamical time needed to completely translate those ribosomes. To measure the mean number of ribosomes on the mRNA, it is necessary to weight each time step by the of the duration of the time step (since each time update by the Gillespie algorithm is chosen from a distribution):

$$\bar{n} = \frac{\sum_j n_j \tau_j}{\sum_j \tau_j}$$

where τ_j is the duration of step j and n_j is the number of ribosomes present at step j.

2.2 Summary of results for the Uniform TASEP with $l = 1$

An analytic solution for particle TASEP has been given by Derrida et al (1992). Three distinct dynamical phases in the current and density were found as a function of α and β. The process features a one dimensional lattice, and particles with hard core exclusion. Particles of size $l = 1$ are either injected into the lattice at the beginning with probability α, terminated with probability β at the end of the lattice, or moved a single lattice site in one direction along the lattice. The uniform case considered does not have hopping rates dependent on position. In table 2.1, J denotes the particle current, ρ_1 denotes the average density at the entrance, ρ_N denotes the average density at the exit, and ρ_{ave} is the average density of ribosomes of the whole system. A phase diagram is shown in figure 2.1

Phase	J	ρ_1	ρ_N	ρ_{ave}
low-density $\alpha \leq \frac{1}{2}$, $\beta > \alpha$	$\alpha(1-\alpha)$	α	$\frac{\alpha(1-\alpha)}{\beta}$	α
high-density $\beta \leq \frac{1}{2}$, $\beta < \alpha$	$\beta(1-\beta)$	$1-\frac{\beta(1-\beta)}{\alpha}$	$1-\beta$	$1-\beta$
coexistence line $\alpha = \beta < \frac{1}{2}$	$\alpha(1-\alpha)$	α	$1-\alpha$	α
maximal current $\alpha \geq \frac{1}{2}$, $\beta \geq \frac{1}{2}$	$\frac{1}{4}$	$1-\frac{1}{4\alpha}$	$\frac{1}{4\beta}$	$\frac{1}{2}$

Table 2.1: Summary of the phases for particle TASEP ($l = 1$).

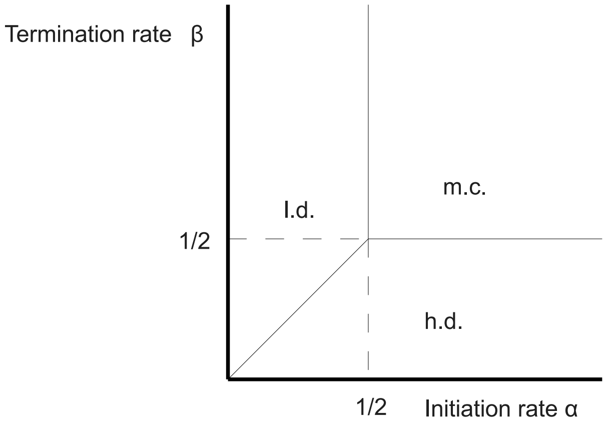


Figure 2.1: Phase diagram for the uniform TASEP model with $l = 1$. The line $\alpha = \beta$ separates the low-density and high-density phases. The system enters the maximal current phase for $\alpha > \frac{1}{2}$ and $\beta > \frac{1}{2}$. For the general case of $l > 1$, the shape of the regions are the same with transition to the maximal current phase now occuring for $\alpha > \frac{1}{\sqrt{l}+1}$ and $\beta > \frac{1}{\sqrt{l}+1}$.

The low density (ld) phase occurs when initiation is the limiting factor. In this case, ribosomes are well spaced out along the mRNA and there are few collisions. This is found to occur when $\alpha < \frac{1}{2}$ and $\alpha < \beta$. The high density phase occurs when the termination rate is the limiting factor, $\beta < \frac{1}{2}$ and $\beta < \alpha$. One finds here a queue of ribosomes waiting to exit the system at a rate determined by β. The transition between the high density and low density phase (through the line $\alpha = \beta$) is a first order transition. There is a discontinuity in the average bulk density when going from one phase to the other. At the maximal current phase the dynamics are governed purely by the exclusion dynamics, so the current and density reaches a constant. This phase occurs when $\alpha > \frac{1}{2}$ and $\beta > \frac{1}{2}$, so that neither initiation and termination limits the dynamics. The transition into the maximal current phase is continuous in density and current.

2.3 Summary of results for the Uniform TASEP with $l > 1$

We have seen in section 2.2 that by adjusting the initiation and termination rate there emerge different solutions in the current and density for a uniform TASEP. The next logical extension to the simple model is to allow for the movement of extended objects, $l > 1$ along the lattice. This is important since each hop of the ribosome is less than its footprint on an mRNA (M. Takanami and G. Zubay, 1964; C. Kang and C.R. Cantor, 1984). Lakatos and Chou (2003), and Shaw et al (2003) investigated the uniform TASEP with particles which are extended objects ($l > 1$) and derived and tested their theoretical solutions with Monte-Carlo simulations. Similar to the $l = 1$ case, the dynamics were also to have three phases.

Uniform TASEP with $l > 1$ has the same phase diagram as that for uniform TASEP with $l = 1$. Instead of the transition to maximal current happening at α^*, $\beta^* = \frac{1}{2}$, it occurs at $\alpha^* = \frac{1}{\sqrt{l}+1}$, and $\beta^* = \frac{1}{\sqrt{l}+1}$ (see figure 2.1). The transition between the high density phase, limited by the termination rate, and the low density phase, limited by the initiation rate still occurs on the line $\alpha = \beta$ and the coverage density which is now, $\rho = nl/L$, is discontinuous when α and β are varied across that line.

The results for the current and density in the three phases are as follows. We have at *maximal current*, the current and average bulk density,

$$J_{mc} = \frac{1}{(\sqrt{l}+1)^2}\,,\; \rho_{mc} = \frac{\sqrt{l}}{\sqrt{l}+1}, \tag{2.1}$$

that is independent of the entrance and exit parameters. This phase gives the maximum ribosome density that the TASEP lattice can have, due to exclusion interactions between the ribosomes, this number is less than 1. The *low-density* stationary phase,

$$J_{ld} = \frac{\alpha(1-\alpha)}{1+(l-1)\alpha}\,,\; \rho_{ld} = \frac{\alpha l}{1+(l-1)\alpha}, \tag{2.2}$$

which is controlled by the initiation rate α, when the termination rate is greater than the entrance rate. And the *high-density* stationary phase,

$$J_{hd} = \frac{\beta(1-\beta)}{1+(l-1)\beta}\,,\; \rho_{hd} = 1-\beta. \tag{2.3}$$

which is controlled by the exit rate β, as ribosomes must queue at the end of the lattice to exit.

The current can be written in terms of the average bulk density in any phase,

$$J = \frac{\rho}{l}\frac{1-\rho}{1-\rho+\rho/l} \tag{2.4}$$

This last equation will appear again in a modified form for our derivation of a TASEP with a single slow site in section 2.5, and it is now worthwhile to denote its origin. The equation gives the current to any TASEP system with a given density of extended objects covering the lattice. It is derived by Shaw et al (2003) for a closed (periodic) system with a conserved number of ribosomes, but it will apply to any TASEP type processes in steady state with a constant number of particles. For open TASEPs where particle number is not conserved, the formula only applies during steady state. The range of validity for the equation in this case is for the coverage density ρ to be between 0 and $\rho_{mc} = \frac{\sqrt{l}}{\sqrt{l}+1}$, the maximal current density, while for a periodic system ρ may go up to 1. Our application of TASEP is the open system, where ribosomes are free to enter and leave the mRNA. The equation can be rearranged in the following manner,

$$J = \frac{\rho_r \rho_h}{\rho_r + \rho_h}$$

where $\rho_r = \frac{N}{L} = \frac{\rho}{l}$, the ribosome number (N) density on a lattice of size L, and $\rho_h = 1 - \rho$ the density of "holes", or space available for ribosomes to hop into. This is a conditional probability, assuming a constant mean density throughout the lattice, for a ribosome (occupying its current site), and to have the next site be a hole (free to hop). Equation 2.4 was confirmed by simulations in Shaw et al (2003), and derived in the context of protein translation by MacDonald et al (1968).

2.4 Comparison of theory for the Uniform TASEP to simulations

Using the Gillespie algorithm, we simulated a TASEP system that allows us to sweep through different values of the ribosome initiation rate. We do this to confirm the validity of the formulae 2.1 to 2.3 for TASEP. In section 1.4.2, we noted that experimental results suggest that the termination of ribosomes from the mRNA is not a rate limiting factor in translation, so that we have set the parameter β to be high relative to the hopping and initiation rate. We simulated uniform TASEPs over a range initiation rates $0 < \alpha < 0.5$, a high termination rate $\beta = 1$, with an mRNA length of $L = 300$ lattice spacings/codons. Results are shown in figures 2.2 and 2.3. Small disagreement between simulation and analytic solution can be explained by the finite size of the lattice. These finite size effects show up particularly when α is close to the transition point between l.d. and m.c. phases ($\alpha^* = \frac{1}{\sqrt{(11)+1}} \approx 0.23$). The derivative of the density with respect to alpha at α^* is smoothed, and the transition between the l.d. and m.c. phases do not show a numerical cusp. For an infinite system, the density enters continously to the m.c. phase with its first derivative discontinuous. Another artifact of finite size effect is the overshooting of the ribosome current ($\approx 4\%$) in the maximal current phase. The shape of the graphs in the low density phase ($\alpha < \frac{1}{\sqrt{11}+1} \approx 0.23$), and the maximal current values agrees with theory. The transition in current is second order in that the current, and its first derivative with respect to α are continuous, while its second derivative is not. These simulations, along with those already done by Shaw et al (2003), and Lakatos and Chou (2003) suggest that simulations reproduce the dynamical behaviours of a TASEP system with extended objects with good accuracy. Theoretical results provides the framework to study our modeling of mRNA translation with TASEP.

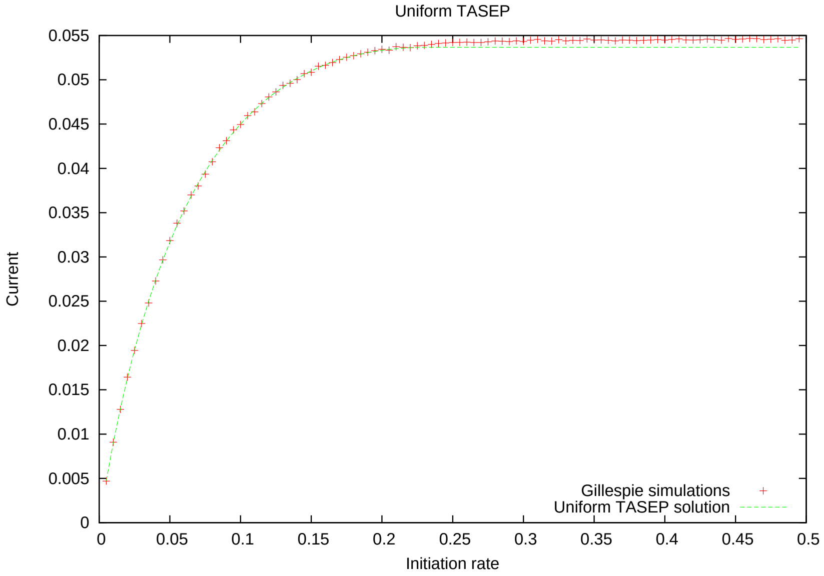


Figure 2.2: This graph shows the current from simulations for uniform TASEP, plotted with the analytical result.

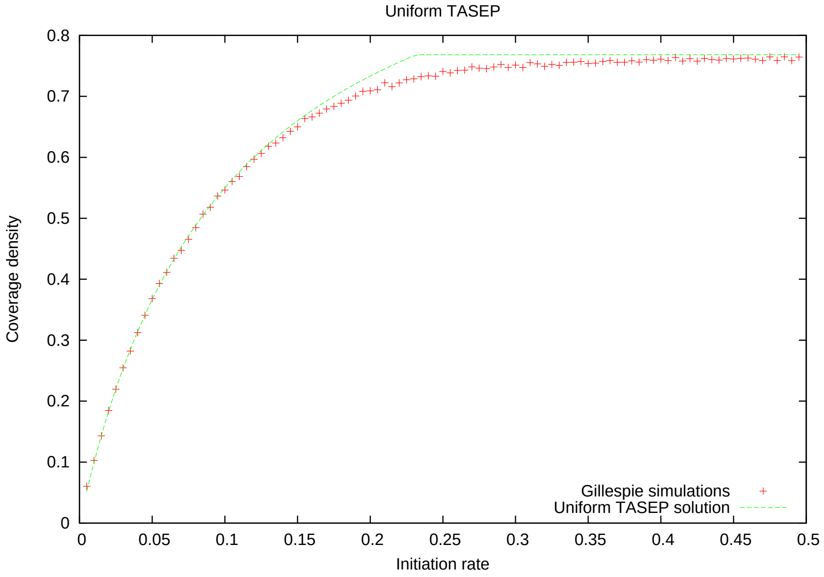


Figure 2.3: This graph shows the average density from simulations for uniform TASEP, plotted with the analytical result.

2.5 Single slow site

When the TASEP lattice has a single slow site, a queue of ribosomes appears behind the blockage. This will help us to investigate queueing in the translation of mRNAs when it is limited by a slow codon on an mRNA. The approach by Shaw et al (2004) will be summarized in this sub-section.

An otherwise uniform TASEP has a slow site placed in the middle of the lattice. The inhomogeneity divided the system in two, and allows for ld/ld, hd/ld, and hd/hd phases. This has a similar phase diagram to the uniform lattice where three phases are allowed. The phase diagram with the corresponding density profiles is shown in figure 2.4.

We derive below the critical initiation and termination rate for which transition between the phases occurs. The mean time for an extended object to pass through an isolated slow site of rate $q < 1$ is,

$$\frac{[1/q + (l-1)]}{l}$$

whence the effective rate is,

$$v_{eff} = \frac{l}{[1/q + (l-1)]} \tag{2.5}$$

The current through the inhomogeneity would be,

$$J_0 = v_{eff} \frac{\rho_{left}}{l} \frac{1-\rho_{right}}{1-\rho_{right}+\rho_{right}/l}.$$

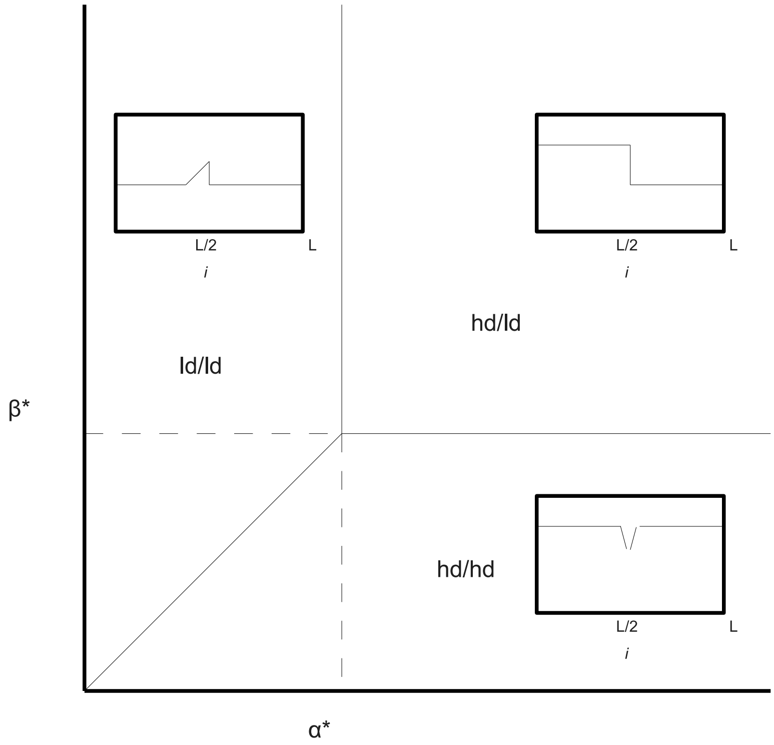


Figure 2.4: This is the phase diagram for TASEP with a single inhomogeneity. Inset are schematic density profile along the lattice where 1 slow site is placed in the middle of the lattice. The three phases are defined by the density in each of the subsystems.

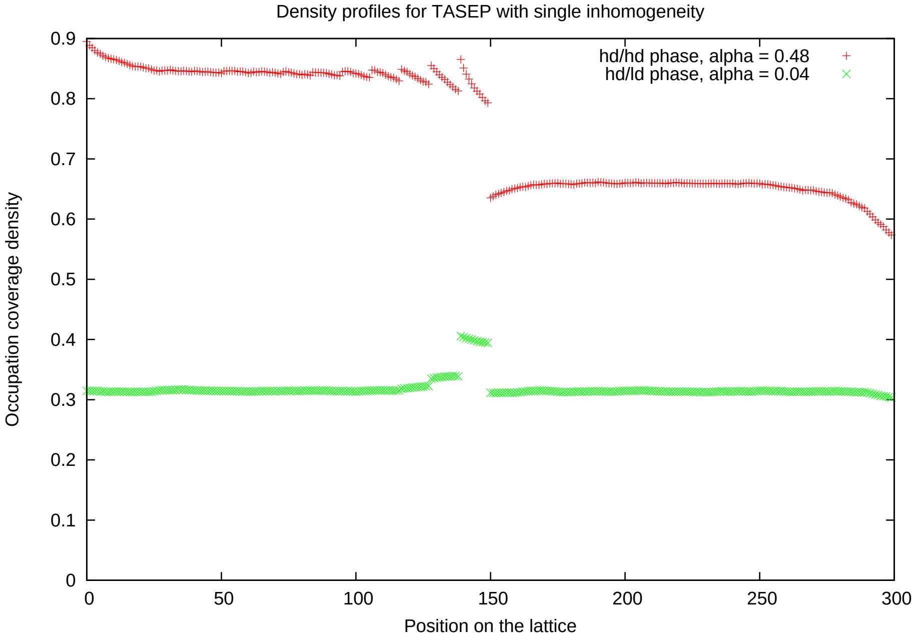


Figure 2.5: This is the phase diagram for TASEP with a single inhomogeneity. Inset are schematic density profile along the lattice where 1 slow site is placed in the middle of the lattice. Periodic structures of period $l = 11$ is observed before the cluster of slow codons, is due to the packing of ribosomes which are of length $l = 11$ against the single slow site. The three phases are defined by the density in each of the subsystems.

This is a modified version of the formula 2.4. It may be interpreted as the conditional probability that there is a ribosome from the left subsystem (its availability is $\frac{\rho_{left}}{l}$, the number density), with a hole from the right subsystem to hop into (the hole-coverage of the right subsystem is $1-\rho_{right}$). This occurs at a rate v_{eff}, the average rate a ribosome passes the site. Using the average rate is allowed in this case since there can only be 1 ribosome at any given time passing the slow site. This is an approximation made in Shaw et al (2004) that takes the mean density of the subsystems to be uniform through the respective subsystems.

In the hd/ld phase, the current from the left subsystem is given by,

$$J_{left} = \frac{\beta_{eff}(1-\beta_{eff})}{1+\beta_{eff}(l-1)} \text{ with density } \rho_{left} = 1-\beta_{eff}$$

which is limited by an effective exit rate β_{eff} from the left lattice. In the right subsystem,

$$J_{right} = \frac{\alpha_{eff}(1-\alpha_{eff})}{1+\alpha_{eff}(l-1)} \text{ with density } \rho_{right} = \frac{l\alpha_{eff}}{1+\alpha_{eff}(l-1)}$$

which is limited by an effective entrance rate α_{eff} into the right lattice.

Solving $J_{left} = J_{right} = J_0$ and noting that the hd/ld can only exist for $\alpha > \beta_{eff}$ and $\beta > \alpha_{eff}$ we obtain,

$$\alpha^* = \beta^* = 1 - \frac{(1+v_{eff})l}{2v_{eff}(l-1)}\left[1-\sqrt{1-\frac{4v_{eff}(l-1)}{l(1+v_{eff})^2}}\right]. \tag{2.6}$$

α^* controls the transition to high density in the 5' lattice, and β^* controls the transition to high density in the 3' lattice. These are the transition values for α, and β for the phase diagram shown in figure 2.4.

By knowing these limits, we can use then use the appropriate extended object TASEP formulae for J and ρ. Here, we will only consider the case where $\beta = 1$, so that only the 5' subsystem can achieve high density, or that the hd/hd phase is excluded from happening. In the ld/ld phase, $\alpha < \alpha^*$, $\beta = 1$ neither subsystem have reached its limiting current/density. The current and density for ribosomes passing the whole mRNA in the ld/ld phase is,

$$J(ld/ld) = \frac{\alpha(1-\alpha)}{1+\alpha(l-1)},\ \rho(ld/ld) = \frac{l\alpha}{1+\alpha(l-1)}$$

which is wholly controlled by the initiation rate. In the hd/ld phase, the current is limited by the inhomogeneity, reaching the limiting current at the critical initiation rate α^* at the left subsystem,

$$J(hd/ld) = \frac{\alpha^*(1-\alpha^*)}{1+\alpha^*(l-1)}.$$

The single slow site causes a spatial jump in ribosome density (*i.e.* a queue begins to form),

$$\begin{aligned}\rho_{left}(hd/ld) =& 1-\alpha^* \text{ and}\\ \rho_{right}(hd/ld) =& v_{eff}\,\rho_{left}(hd/ld)\end{aligned}$$

since we know that $v_{eff} < 1$. This gives,

$$\begin{aligned}\rho(hd/ld) &= \frac{L_{block}}{L}\rho_{left}(hd/ld) + (1 - \frac{L_{block}}{L})\rho_{right}(hd/ld)\\ &= \frac{L_{block}}{L}(1-\alpha^{*}) + (1 - \frac{L_{block}}{L})v_{eff}(1-\alpha^{*})\end{aligned}$$

where L_{block} denotes the position of the single slow site along the lattice. The density contribution from each subsystem depends on the lattice fraction that the respective subsystems occupies.

We simulated a system with a single slow site. Results are shown in figures 2.6 and 2.7. The system transitions into the hd/ld phase earlier than the transition for the uniform TASEP into the maximal current phase. The current profile for the mRNA with a blockage shows close agreement with the theory except at the transition. This is again due to finite size effect of the lattice length. The density profile shows strong disagreement at the transition due to finite size effect and the 'softness' of the blockage.

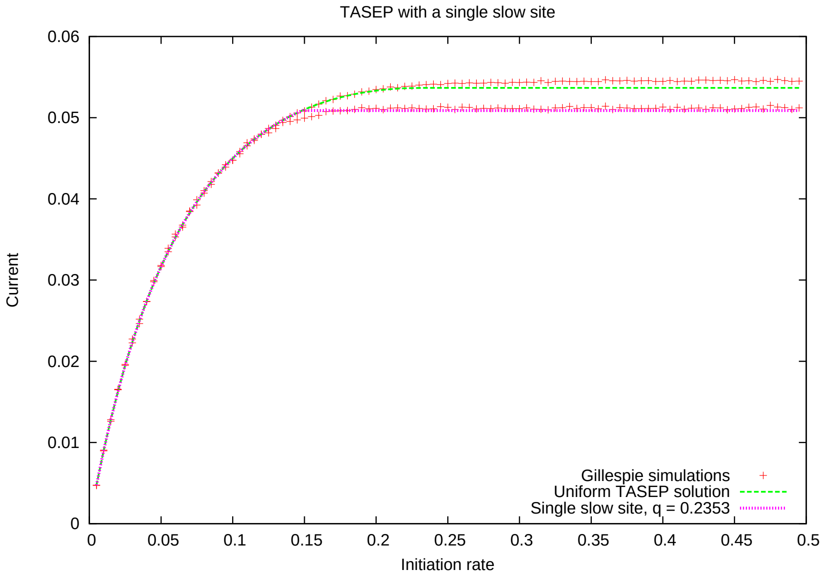


Figure 2.6: We see a transition to the high-density phase sooner when the slow site is slower. Plotted are the currents from simulations for two different rates for the cluster and their analytical results.

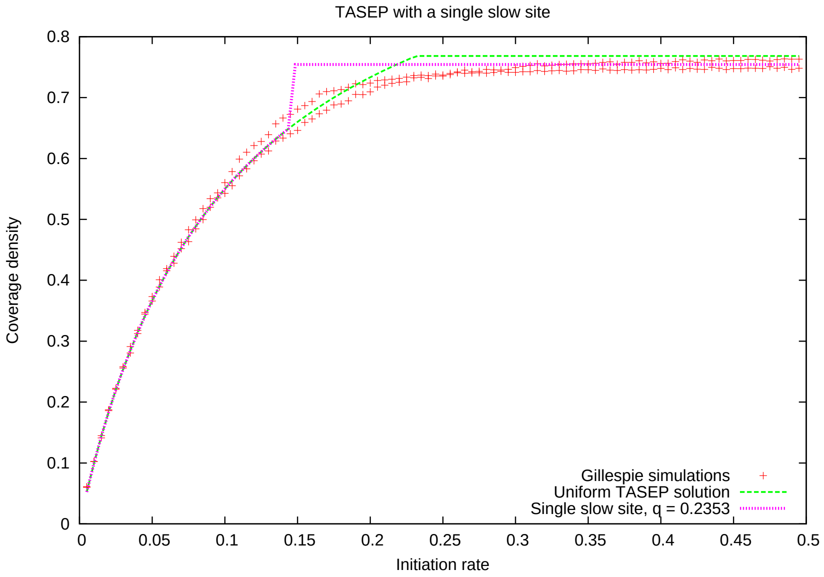


Figure 2.7: We see a transition to the high-density phase sooner when the slow site is slower. Plotted are the average density from simulations for two different rates for the cluster and their analytical results.

2.5.1 Slow sequence of length less than l

Here I give an extension to (Shaw, Kolomeisky, Lee, 2004), and derive an approximate solution to a small sequence of slow codons located in the bulk of a TASEP lattice. This is motivated by the fact that the slowest part of an mRNA will cause ribosomes to queue. So that, the slowest small sequence of codons on an mRNA may determine translation behaviour.

For a slow sequence of length $s \leq l$, we use an effective rate similar to the single defect case to approximate the inhomogeneity. Here we replace the definition of v_{eff}, equation (2.5) above with,

$$v_{eff} = \frac{l}{[s/q + (l - s)]}$$

the effective rate for which a ribosome passes s slow sites. As the slow sequence becomes larger than l, it can accomodate more than 1 ribosome, but with interactions, the approximation breaks down. The effective rate should be a good approximation for s less than l, as the denominator in v_{eff} may become negative.

If there is a region of l consecutive slow codons, each with $v = q < 1$, then the effective velocity in this region is $v_{eff} = q$. The derivation in section 2.5 applies with $v_{eff} = q$. To test this, we simulated a system with 11 slow codons placed in the middle of an otherwise uniform TASEP lattice. The termination rate is again put to be $\beta = 1$ and the lattice is $L = 300$ lattice spacings or codons. Results are shown in figures 2.8 and 2.9. As with before, at the transition, the simulations disagree with analytical results due to finite size effect, but as the strength of the blockage is increased the 'jump' is better approximated.

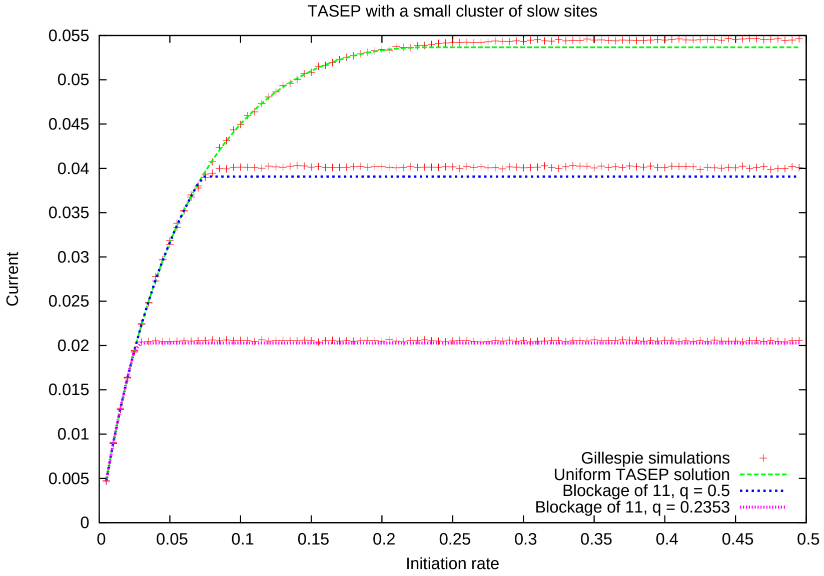


Figure 2.8: We see a transition to the high-density phase sooner when the cluster is slower. Plotted are the currents from simulations for three different rates for the cluster and their analytical results.

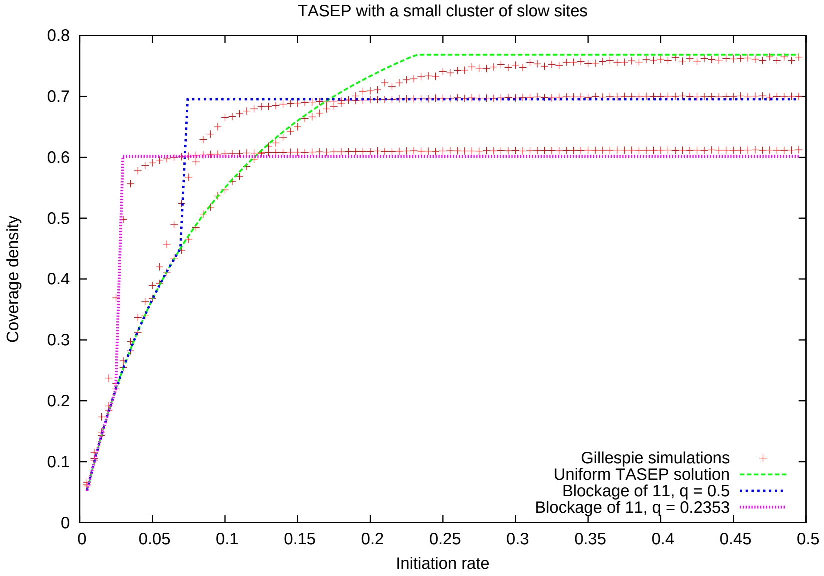


Figure 2.9: We see a transition to the high-density phase sooner when the cluster is slower. Plotted are the average density from simulations for three different rates for the cluster and their analytical results.

2.6 Ramp of slow codons

Tuller et al (2010) observed, averaging over many genomes, that rare codons are used at the beginning of genes. This lead to the conclusion that a 'ramp' of slow codons at the beginning of genes is an evolutionarily conserved translational mechanism. The position of these slow codons at the beginning of genes would cause interference to the initiation process, effectively lowering it. The lowered initiation rate causes a decrease in ribosomes on the mRNA and thus increase distance between ribosomes. When ribosomes are spaced out the chance of queueing later in the sequence is reduced. This means that the position of a cluster of slow codons causes a different effect other than queueing when it is placed at the beginning of the sequence. Quanlitative differences for particle TASEP ($l = 1$), between a bottleneck of slow codon near the edge, and in the bulk were observed by Greulich and Schadschneider (2008). In this case, they saw that the current J do not have a transition into the maximal current phase, but shows a 'softened crossover'.

As seen in the previous section, a short ($\leq l$) sequence of slow codons on an otherwise uniform TASEP causes significant queueing of ribosomes. In this section, we will explore what happens when the cluster of slow codons is placed at the beginning of the gene. We placed a cluster of size $l = 11$ starting at the site $x = 1$ and ending at $x = l = 11$ and assigned $v_{ramp} < 1$ as the rate of these slow codons. The rate of the TASEP after the ramp is denoted v_{rest} and was set equal to 1.0. Doing so allows us to test ramped TASEP with different $\frac{v_{ramp}}{v_{rest}}$.

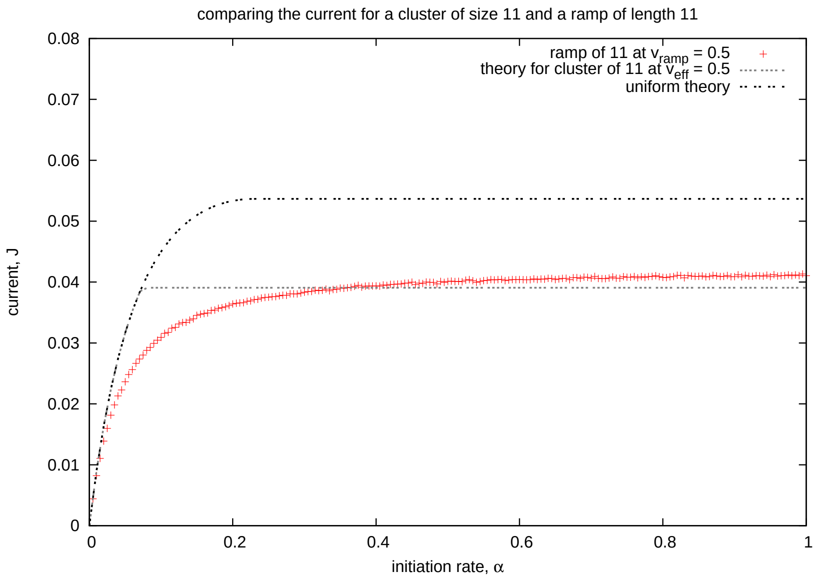


Figure 2.10: Simulations for $v_{ramp} = 0.5$ and $v_{rest} = 1.0$. The effect cluster rate is the same the v_{ramp} and both sequence of slow codons is of size $l = 11$.

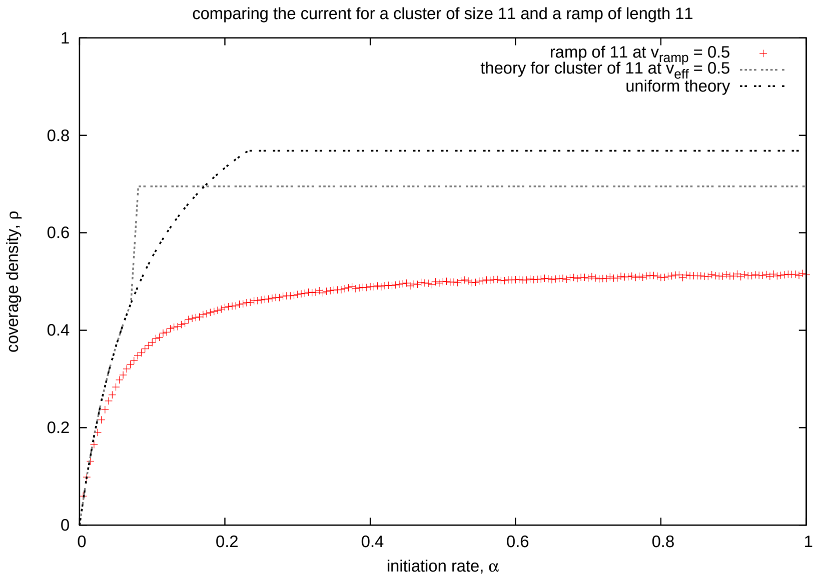


Figure 2.11: Simulations for $v_{ramp} - 0.5$ and $v_{rest} = 1.0$. The effect cluster rate is the same the v_{ramp} and both sequence of slow codons is of size $l = 11$.

The behaviour of a block at the beginning of a sequence is qualitatively different from that of cluster and the uniform TASEP. Ribosome density in the ramp region increases compared to uniform TASEP because the initiation rate compared to v_{ramp} is high and ribosomes are added as soon as there is enough space. Although the density of the ramp region is high, the current (which scales with v_{ramp}) from the region is slowed by the usage of slow codons. Both the current and the density are lowered for the whole system, due to the ramp limiting the number of ribosome which passes onto the rest of the sequence per unit time, and that this spaces out the ribosomes. The current for the ramped system exceeds the cluster limiting current at high initiation rate and was not expected. The current is allowed to reach a value greater than half of the uniform TASEP solution since a ramp of size $l = 11$ is just small enough that only 1 ribosome may occupy any of the sites with $v_{ramp} < 1$ at a given time. Due to a lack of ribosome-ribosome interaction, the current from the ramp may exceed $\frac{v_{ramp}}{\sqrt{11}+1}$. Though at the same time, a smaller effective initiaiton rate is imposed on the rest of the system since the current throughout the lattice must be equal during steady state. Since no high density queues were formed, the ramped system maintains a lower ribosome density at all initiation rates. A variety of v_{ramp} were tested and the results are shown in figures 2.12, and 2.13.

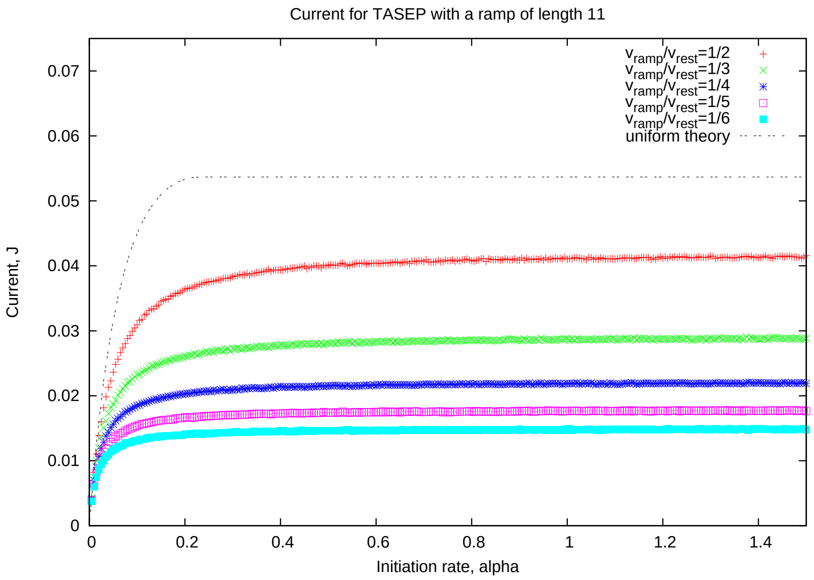


Figure 2.12: Simulations for six different ramp rate v_{ramp} with $v_{rest} = 1$. The current is lower for smaller v_{ramp}.

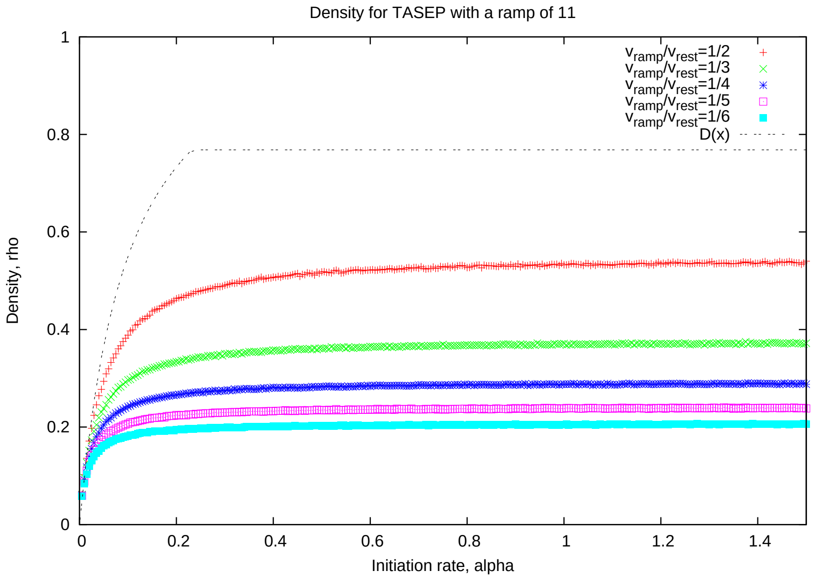


Figure 2.13: Simulations for six different ramp rate v_{ramp} with $v_{rest} = 1$. The density is lower for smaller v_{ramp}, *i.e.* ribosomes space out.

From these figures, we see a gradual decrease in current and density as the rate of the ramp is decreased, as expected. A similar analysis to section 2.5 can be done with a ramped TASEP, where there are two subsystems: the ramp, and the rest of the lattice. These can be considered as two TASEPs which are coupled to each other, and the current of ribosomes exiting from the ramp defines the effective initiation rate into the rest. We see that as the ramp approach the length of the whole lattice, solution would simply be given by $J_{ramp}(\alpha) = J(\frac{\alpha}{ramp})$. So that the effective initiation rate scales α from the end of the lattice by v_{ramp}. However, we find that the functional form for the solution to the current of a system with a short ramp (of length $\approx l$) is different from that of the uniform system. Due to the finite size of the ramp, the normal TASEP solution cannot be considered. The main effect of the ramp is to limit initiation by scaling initiation by roughly v_{ramp}, and limits the current by a factor of $\frac{v_{ramp}}{v_{rest}}$.

Chapter 3

Using the TASEP model to simulate translation of real gene sequences

3.1 Assigning translation rates to codons

According to section 1.5, there are various schemes with which the adaptation of codons can be measured. Our model can test the role codon adaptation plays in the translational mechanism. As discussed in section 1.3, the usage of frequent codons, and tRNA levels are correlated with growth rate of an organism, and that these measures of adaptation have co-evolved. Since growth rate and expression level have been identified with these measures, it is suggestive that codon bias and tRNA usage determine the translational efficiency of mRNAs. Exactly how translational efficiency works in a kinetic model will be investigated in this section. In our simple kinetic model of translation, codons are processed by the ribosome at different rates. We will assign codon rates according to previous evidence from experiments and simulations, guided by bioinformatical measures of codon usage and tRNA levels.

3.2 Likelihood signal of translational selection

Higgs and Ran (2008) has conducted a detailed study on the relationship between tRNA recognition and translational efficiency. Further work is carried out by the same authors in 2012, testing the statistical significance of translational selection at the codon level between high and low expression genes. They defined the following quantity to measure the statistical significance of a selection scheme,

$$\delta_i = \ln\left(\frac{\phi_i^H}{\phi_i^0}\right). \tag{3.1}$$

Where ϕ_i^H is the relative frequency of codon i for an amino acid in a high expression set (normally the ribosomal proteins, and elongation factors), and ϕ_i^0 to be the genome wide frequency for those codons sharing the same amino acid. When the log value of a codon is negative it indicates that it is selected against in high expression genes. In our model we take this as indication that a codon has slow translation rate. The frequencies, and δ values of codons are shown in table A.3.

Taking the average of δ's for codons in a gene, we can compute the adaptation for a particular gene. The expression,

$$\delta_{gene} = \frac{\sum_k \delta_{i_k}}{L_{gene}} \tag{3.2}$$

averages δ_i for an amino acid i_k used at site k in a gene of length L_{gene}.

3.3 Experimentally fitted rates

Mitaria et al (2008) carried out simulations of a version of a TASEP model and used these to fit data from experiments (mentioned in section 1.4.1). The model they used was similar to the one in this thesis, with the modification that mRNAs are allowed to decay, and the simulation keeps track of the radioactive signal from elongating proteins. The ribosome was allowed to pick up a radioactive signal as it passes a methionine codon, and the duration or total amount of radioactive methionine allowed was determined by the duration of an initial pulse of the label. After a predetermined number of the translations, or ribosomes that have hopped to the end, the mRNA is stopped from initiating new ribosomes, indicating decay of the mRNA. The mRNA is parameterized by the ribosome initiation rate and codons' elongation rates, and the number of translations before initiation is stopped. A range of parameters were simulated and the data which most closely fit the experimental data from a previous experiment were reported. Three categories of codons were required to fit the radioactivity data. The codon categories were determined using the codon adaptation index (defined in section 1.5.1). The translation rates of the categories which best fit the data were found to be 35.0, 8.0, and 4.5 codons per second and agree with the rates for specific codons found experimentally in Sørensen and Pedersen (1991). This used an intiation frequency of 0.9 ribosomes per second and they estimated that the mRNA gets translated 40 times before it decays.

The assignment of codon speed categories presented by Mitarai et al (2008) were based on the codon adaptation index (see section 1.5.1), which weights the selection of each codon using only the frequencies in the reference set of high expression genes.

The measure δ used by Higgs and Ran depends on the frequency in the high expression set and the whole genome. The rate categories used by Mitarai et al (2008), and values of δ are shown in table A.3.

3.4 Local tAI

Local tAI is defined by Tuller et al (2010), and the index is used to determine the translation rate of codons. Using equation 1.1, a local measure of codon adaptation to the tRNA pool is obtained. Simulations were conducted by Tuller et al setting the translation rate of each codon to be its local tAI value. Examples of local tAI for *E. coli K12* are shown in table A.3.

Since the most abundant tRNA gene in *E. coli K12* translates the AUG codon (amino acid: methionine), it is assigned a value of 8 for the unnormalized tAI ($W_{max} \times tAI$. This corresponds to 8 copies of the gene. Since the wobble position of the AUG codon matches its sole anti-codon in an energetically favourable Watson-Crick pair (G-C) the selective contraint b_{ij} for the pairing has a value of 1. In the case of leucine, there is one copy of each UUA, CUC, CUA anticodon pairing tRNA gene that only forms Watson-Crick pairs (A-U,C-G,A-U, respectively) then these codons are assigned an adaptiveness of 1. In the case of the UUA codon for leucine, there are 2 copies of the tRNA gene that only form a Watson-Crick pairing, then this codon is twice as well adapted as the UUA, CUC, and CUA codons. Since codon UUG may pair with anti-codon CAA, and UAA, then its adaptiveness is more than UUA, but less than twice the value because the pairing strength of UUG-UAA is less than UUG-CAA (the G-U pairing is less efficient than G-C pairing). Because there are no copies of the anti-codon AAG tRNA gene that means CUU can only form a less than optimal pairing with UUU, so that its local tAI is a small number.

The other values in table 1.1 are to show the relative value of local tAI between select codons. In their model, Tuller et al (2010) used local tAI values as the hopping rate for each codon.

3.5 Simulations and relevant features to translational selection

3.5.1 Parameterizing elongation rates and selection of genes from *Escherichia coli*

In this section we test our model of translation with real sequence from the genome of *Escherichia coli* using the estimate of the translation rates as prescribed in table A.3. We decided that we would assign codon rates similar to that done in (Mitarai et al 2008). We opted against using tAI because the values were dimensionless numbers, and did not have an estimate for the initiation rate. Also, the results of Tuller et al (2010) had different predictions for relative codon rates than the experimentally fitted elongation rates by Mitarai et al (2008), and largely disagreed with our δ measure of codon usage in the *E. coli* genome. For example, in the amino acid family valine (Val), the codon GUU has the lowest tAI (0.878) but shows the strongest selection in the family with $\delta = 0.697$. In the same family (Val), the codon GUC is assigned a negative δ value but the second largest tAI in the family. For threonine

(Thr), the codon with the smallest tAI is assigned the largest δ. Disparity between tAI as defined by dos Reis et al (2005) and real codon rates is exacerbated when the codon-anticodon pairing constraints (b_{ij} in equation 1.1) are extra parameters that are obtained by recursively maximizing the tAI values and expression levels but not on experimental measurements. The basic model to the pairing constraints also did not take into account the wobble pairing of U-U nucleotides and modified basis which was found to be important in predicting selection (Ran, and Higgs. 2010). We chose to use the codon rate scheme from the Mitarai et al (2008) result because with few modifications, the codon rates category assignment agrees with our δ values.

The findings of Mitarai et al (2008) were modified slightly so that it becomes more consistent with the δ value for the following codons: CUG, AUU, GUG, UAA, UAG, and UGA. Since CUA for the amino acid leucine (Leu) had a small δ, its rate cannot be more than the other synonymous codons, so it is assigned the category of C. AUU for isoleucine (Ile), has a negative codon expression, but was assigned a rate category of A is inconsistent since most A codons had positive expression values ($\delta > 0$), and a synonymous codon that is also assigned a rate category of A had positive expression. Similarly for codon GUG of valine, that has δ similar to that of synonymous codon GUC. The remaining codons, UAA, UAG, and UGA are stop codons, and were assigned the median codon rate.

Using the modified scheme to assign codons rates, we would expect the average elongation rate have a positive correlation with the adaptation δ. The average elongation rate is calculated by equation 3.3. Figure 3.1 shows the relationship of δ_{gene}

and the gene's average translation speed as set by the three categories.

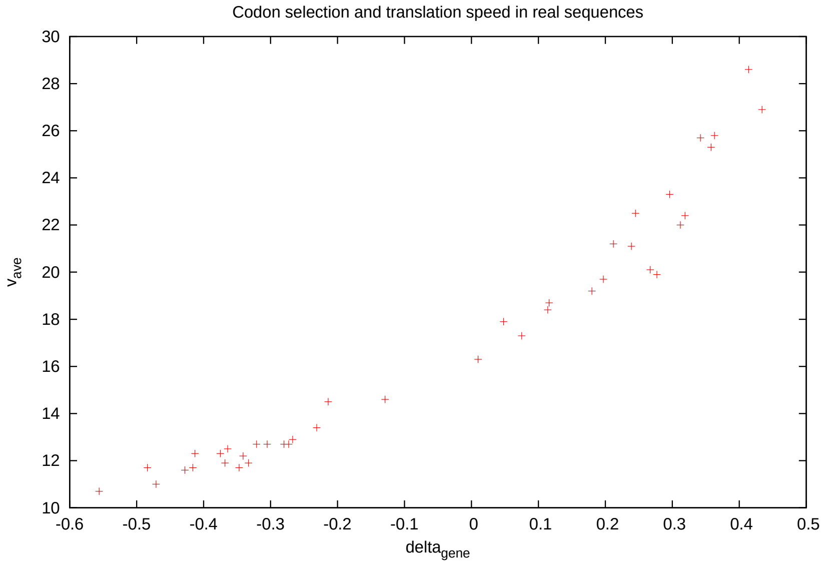


Figure 3.1: δ was calculated for genes, and is plotted as a function of average cond rate. More rapidly translated mRNAs uses more frequent codons; plot shows positive correlation, as expected. Each point on the graph represents a gene.

We have chosen genes to simulate from the genome of *Escherichia coli* that represents a spectrum of parameters in our model. The complete table of genes simulated are shown in the appendix (A.5). The δ measure for the selection ranges from -0.556 to 0.434. The ribosomal proteins are labelled L and S for large and small subunit, and they were used as the high expression reference genes. The other genes with high values of δ are for often used enzymes, and the ones which have low δ were randomly selected from the genomes.

We investigated a total of 4262 genes from the *E. coli* K-12 genome calculated the gene value of δ. 228 of these genes have $\delta > 0.0$, 1524 of these genes have a value $0.0 \geq \delta > -0.3$, and 2510 had $\delta \leq -0.3$. Most genes in the genome had $\delta \leq 0$. The reference set comprise most of the genes we investigated using the TASEP model. Figures 3.2 and 3.3 plots sets of genes with different values of δ, the average adaptation measure δ, and codon rate at each position starting from the beginning of genes. The plots shows that the average δ, and codon rates are higher for high expression genes, than that for intermediate, and low expression genes. We observed, as in (Tuller et al, 2010), that there is a drop in codon adaptation at the beginning of genes, and we would like to investigate its role in the kinetics of translation.

3.5.2 Uniform, cluster, and ramp approximations

From our theoretical exploration of TASEP in chapter 2, we observed that the initiation rate, and the elongation rate of codons to be important in determining the dynamics of translation. We approximated real sequences using their average translation velocity and the position and strength of a cluster, and tested their similarity to the dynamics of translation of real codon sequences. We attempt to extract the salient features of real sequences in translation for our kinetic model. Select sequences are taken from the genome, and we measured the current and density produced.

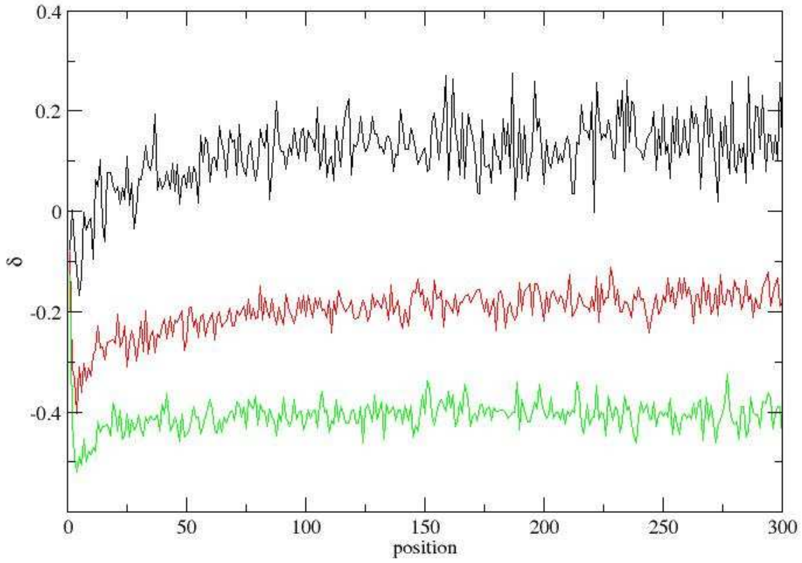


Figure 3.2: The genome average of the codon $/delta$ values as a function of position for all genes aligned at their start codon. The three color curves denotes ranges of δ_{gene} sampled. The ranges are for black $\delta > 0$, red $-0.3 < \delta < 0$, and green $\delta < -0.3$.

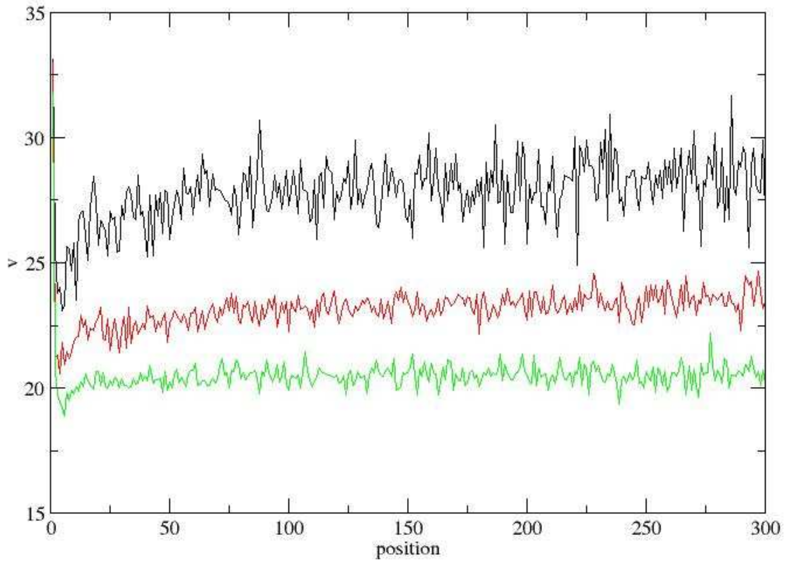


Figure 3.3: The genome average of codon speed based on the three categories. Averages are taken as a function of position for all genes aligned at their start codons. The three color curves denotes ranges of δ_{gene} sampled. The ranges are for black $\delta > 0$, red $-0.3 < \delta < 0$, and green $\delta < -0.3$.

Uniform approximation

First, we approximated the sequences with a uniform TASEP that translates at the average rates of the codons that comprise it. The average translation rate was computed by averaging the mean time needed to translate each codon on the sequence.

$$v_{ave} = \frac{1}{\tau_{ave}} = \frac{L}{\sum_{x=1}^{L} \frac{1}{v_x}} \tag{3.3}$$

Let $J(\alpha)$ be the solution for the uniform TASEP with $v_{ave} = 1$ (see equations 2.1 to 2.3). Let $J(v_{ave}, A)$ be the solution for a uniform TASEP with mean velocity v_{ave} and initiation rate A. It follows that

$$J_(v_{ave}, A) = v_{ave} J(\alpha = \frac{A}{v_{ave}}) = \frac{A(1 - \frac{A}{v_{ave}})}{1 + (l-1)\frac{A}{v_{ave}}}$$

so that $A = v_{ave}\alpha$ is the absolute initiation rate (in units of ribosomes per second). The maximal current phase is reached at,

$$A_{mc} = \frac{v_{ave}}{\sqrt{l}+1} \text{with current, } J(v_{ave}, A_{mc}) = \frac{v_{ave}}{(\sqrt{l}+1)^2}$$

The quickest possible mRNA using our assignment of codons rates is at 35.0 codons per second, and this would correspond to a transition to the maximal current phase at $A_{mc} = \frac{35.0}{\sqrt{11}+1} = 8.1$ ribosomes per second with a current of $\frac{35.0}{18.6} = 1.88$.

At low initiation rate the ribosomes will be widely separated, and no queues would form. The time taken for one ribosome to complete the full mRNA sequence should be the same as the average uniform sequence. Therefore at low initiation rate the

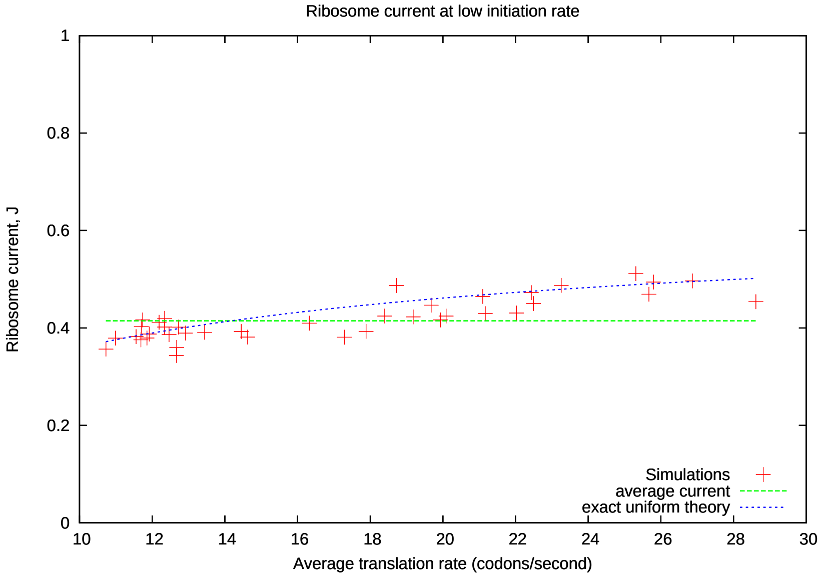


Figure 3.4: The ribosome current of the real sequence at low initiaton rate. A plot of the uniform theory at constant initiation rate ($A = 0.625$) is shown as a function of the average codon speed. Points are scattered around the curve, indicating that the uniform TASEP theory is a good approximation for mRNA translation at low initiation rate. The average current is also computed, this corresponds to the average on-rate of ribosomes. The data points are sampled for $A = 0.625$ ribosomes per second and each point on the graph represents a gene.

current in the real sequence should be close to that in a uniform TASEP with the same v_{ave}. Figure 3.4 compares real sequences at low initiation rate with their average velocity. We observe that there is some agreement between the uniform theory at low initiation rate and the current from real sequences. The uniform approximation and real sequence agree at low initiation rate because collisions are important, but effects from other features, such as a cluster have not emerged. There are simply not enough ribosomes on the mRNA to allow significant queueing.

Experiments on the lacZ gene from *Escherichia coli* estimate a ribosome loading frequency of 1 per 2.2 seconds (Kennell, and Riezman. 1977) and 2.8-3.8 seconds (Sørensen, and Pedersen. 1991), and simulations with semi-stochastic ribosome initiation give 1 per 2.3 seconds (Mitarai et al. 2008). The average loading time is the average time between ribosomes on an mRNA, the loading rate is the inverse of that. The rate at which ribosomes complete a translation of an mRNA is the inverse time between ribosomes completing a translation. At steady state the average number of ribosomes binding onto the lattice per unit time should be the same as that exiting to keep the number of ribosomes constant. This means that the current of ribosomes from an mRNA equals the loading rate. The initiation rate that parameterizes our simulation is not the loading rate of ribosomes, since it does not take into account ribosomes still at the beginning of mRNAs that can block initiation. The absolute kinetic rate over estimates the frequency. The average current over the genes investigated gives a loading rate of $\approx \frac{1}{0.4} = 2.5$ seconds, and is shown in figure 3.4.

Cluster approximation

At high initiation rate, we expect that the system reaches a limiting current defined by the slowest codons on the mRNAs. Although real genes may contain any sequence of codons, we approximated clustering behaviour on real sequences by finding the slowest codons and simulating it using the TASEP model. To find the slowest cluster, we scanned sequences averaging the rate of blocks of $l = 11$ codons, recording the slowest block speed and its position. The rate of the slowest l cluster for a gene of length L is defined,

$$v_{block} = \min_{1 \leq k \leq L-l} \frac{l}{\sum_{x=k}^{x=k+l} \frac{1}{v_x}}$$

and the rate outside the blockage $v_o$is defined,

$$\begin{aligned} \tau_o = \frac{L-l}{v_o} &= \tau_{ave} - \tau_{block} && (3.4) \\ &= \frac{L}{v_{ave}} - \frac{l}{v_{block}} && (3.5) \end{aligned}$$

We approximated the real sequence by placing an l block of codons at rate v_{block} on a lattice with rate v_o. The position of the block is where the l block with average rate v_{block} occurred on the real sequence. A plot of the average rate of the mRNA, and the rate of the slowest $l = 11$ cluster is seen in figure 3.5. The use of faster codons on average also leads to a higher rate for the slowest cluster, this shows that slow codons do not have a particular affinity to be close together in high expression genes. The cluster causes the system to queue when the initiation rate of new ribosomes is sufficiently high. When the system is queueing, it is in the hd/ld region of the TASEP with a blockage phase diagram. The critical initiation rate is defined to be the initiation rate for when the system transitions into the hd/ld phase.

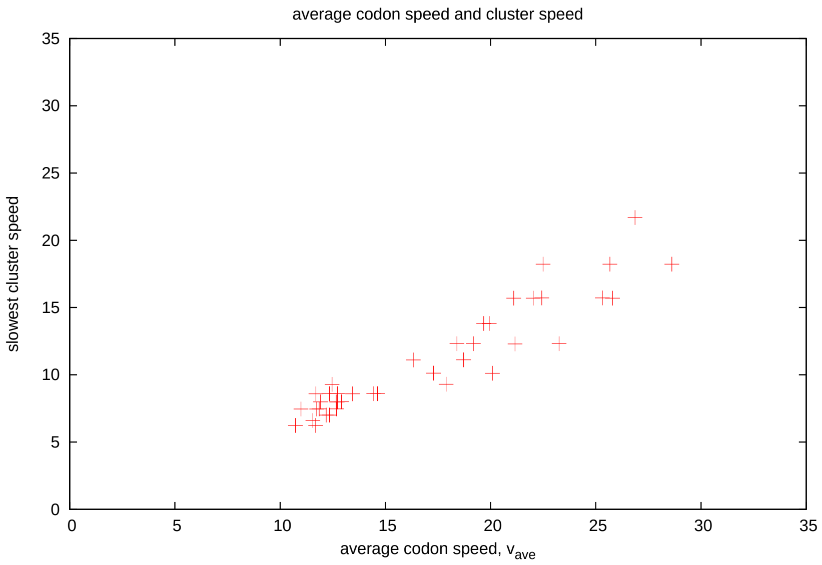


Figure 3.5: This is a plot of the slowest cluster rate on a gene against the average rate of the gene. There is a positive correlation between faster genes that uses faster codons with the rate of its slowest cluster.

The equation 2.6 is applied to find the critical intiation rate for when queueing begins. The absolute initiation rate, in units of ribosomes per second is defined by,

$$A^* = v_0 \, \alpha^*(v_{eff}) \tag{3.6}$$

where v_{eff} is now defined by,

$$v_{eff} = v_{block}/v_o$$

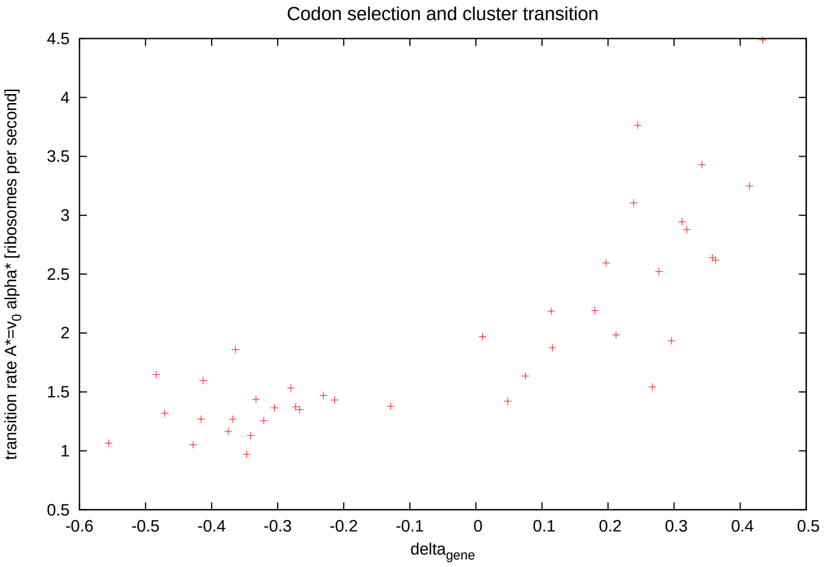


Figure 3.6: This graph plots the critial initiation rate for the cluster approximation of real sequence. At the critical initiation rate $A^* = v_o\alpha^*$ the system transitions into the hd/ld (‘queueing’) phase for a TASEP with a cluster. A positive correlation is observed for high expression genes as measured by δ_{gene}. The limiting current for the approximated mRNA is reached when the initiation rate is greater than A^*. Each point on the graph represents a gene.

We see in figure 3.6 that the critical initiation rate is some-what positively correlated with genes that had $\delta > 0$. This means that for a given initiation parameter, the genes with high δ would queue later, and therefore would be more robust against queueing. It is to be expected that genes with higher v_{ave} would be in the queueing phase at high A, unless there were particularly slow clusters of codons in these sequences. This means that blockages are less severe as they occur at higher initiation rate. The low δ genes does not show the same positive correlation, and there is no selection for avoidance of queueing in these sequences.

3.5.3 Comparison of the cluster and uniform approximations to the real sequence

In this section we will compare the approximations, and evaluate their validity in approximating real sequences. Since real sequences appear well approximated by the uniform TASEP when the initiation rate is low (see section 3.5.2, figure 3.4), we will consider the approximations under high initiation. We performed simulations at an initiation rate $A > A^*$ to allow the effect of queueing.

Figures 3.7 and 3.8 compares at high initiation rate the uniform and cluster approximations. At high initiation the cluster approximation out-performs the uniform approximation, so that queueing is significant. The uniform approximation tends to over estimate the number of ribosome on the sequence, while the cluster approximation has a closer approximation to the ribosome profile. We attribute this over estimation to the fact that a cluster may appear anywhere along the sequence, and their correct positioning is an extra parameter by which a real sequence is approximated. The position of the slow codons defines the length of sequence in the high density phase and the number of ribosomes in the queue. The configuration of codons on an mRNA becomes important when the initiation is high, or that the initiation induces a queueing phase on the mRNA.

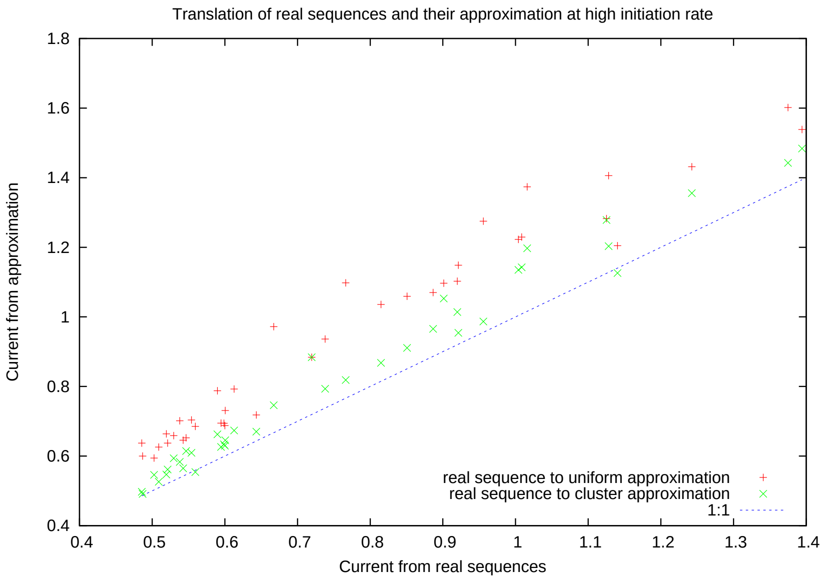


Figure 3.7: We compare the uniform approximation and the cluster approximation to the ribosome current of the current with the real sequence with high initiation rate $A = 35.0$. The cluster approximation points lie more closely to the 1:1 line, and better approximates the real sequence. Each point on the graph represents a gene.

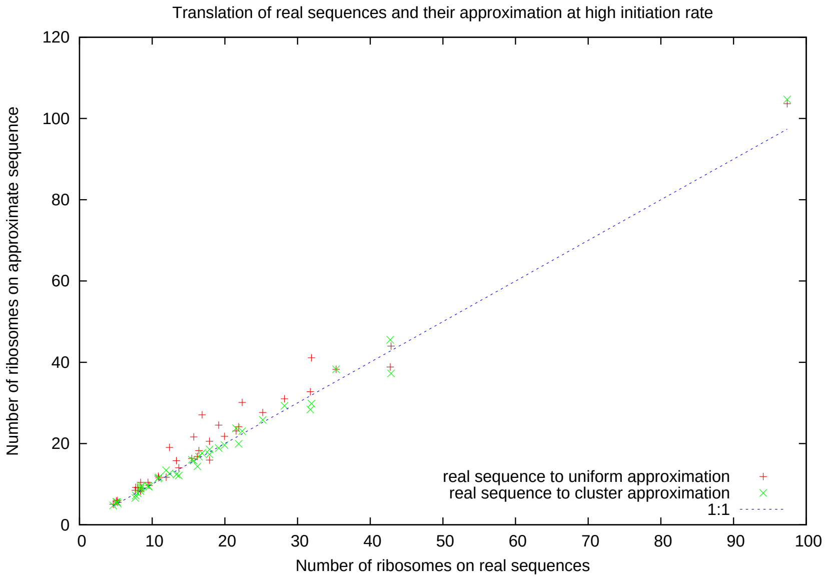


Figure 3.8: This plots the ribosome density achieved using the approximations and comparing it to the real sequence's ribosome density. Data were taken from simulations running with high initiation rate ($A = 35.0$). Each point on the graph represents a gene.

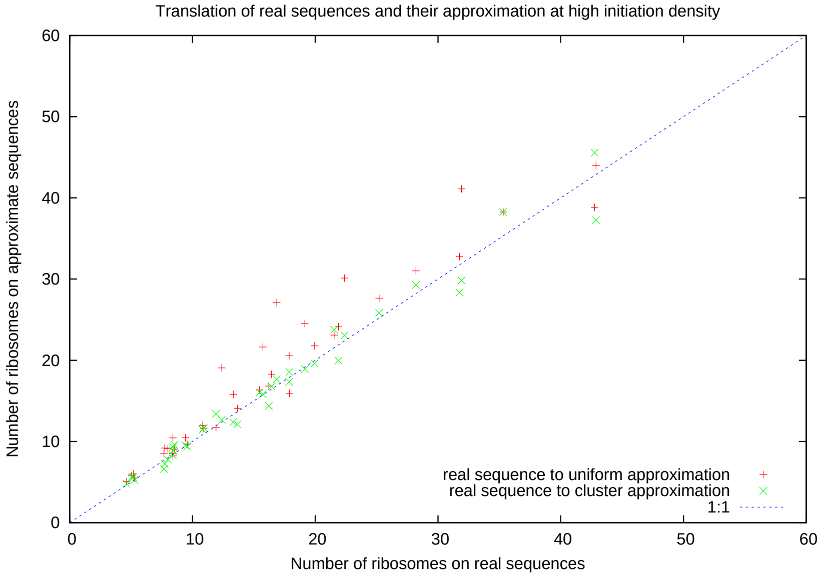


Figure 3.9: For a more careful comparison of the approximations in the number of ribosomes a point which is far from the scale is removed from the graph 3.9. Each point on the graph represents a gene.

The heterogeneity of gene sequences leads to many possible profiles of ribosome density on an mRNA. The codon speed profile of the sequence for enoyl-acyl carrier protein reductase, its cluster approximation, and ribosome density at two initiation rates is shown in figure 3.11. There is an anticorrelation between the sliding window average and the density, that is, the faster the codon the less time a ribosome spends there. We compare the real sequence density with the approximations at low and high initiation rate in figures 3.12, 3.13, 3.14, and 3.13. The dip in ribosome density near the end of the mRNA is an artifact of the ribosome counting method, where a ribosome is considered to have exited when the site closest to the end hop from the last site. As said before, features in the codon arrangement is better approximated at high initiation rate by the cluster approximation, where for low initiation the real sequence

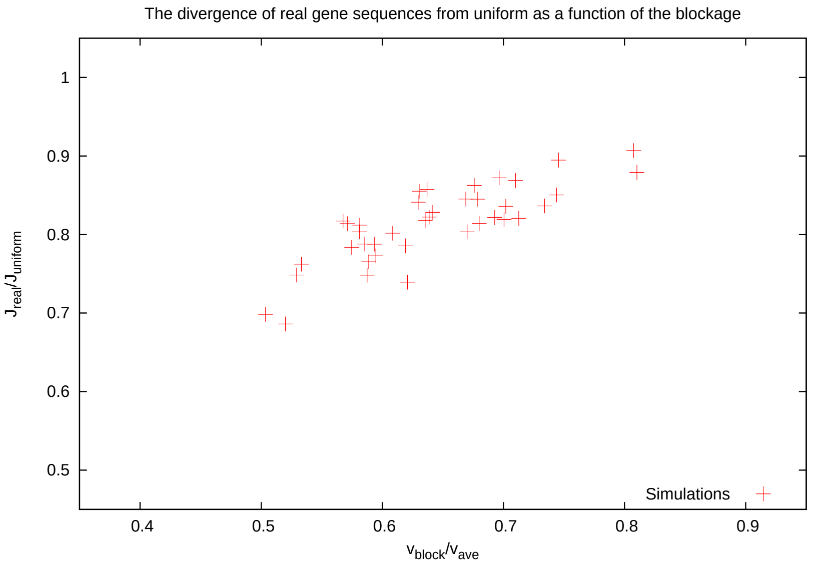


Figure 3.10: The difference between a real sequence and the uniform approximation is plotted as a function of blockage strength. The blockage is translated more quickly the further along on the x-axis, and the y-axis approaches 1 when there is no difference between the current from the real sequence and the uniform approximation. There is a positive correlation between relative cluster rate and how close the current is approximated by the uniform approximation at high initiation. Data points are taken at $A = 35.0$.

profile fluctuates around the uniform approximation. Furthermore, we see that as the slowest cluster of any sequence approach the average rate, making any significant blocks in codon usage along the mRNA less apparent, the real sequences become better approximated by the uniform sequence. This is shown in figure 3.10, where the simulated real sequence approach the uniform approximation as the blockage rate is closer to the average.

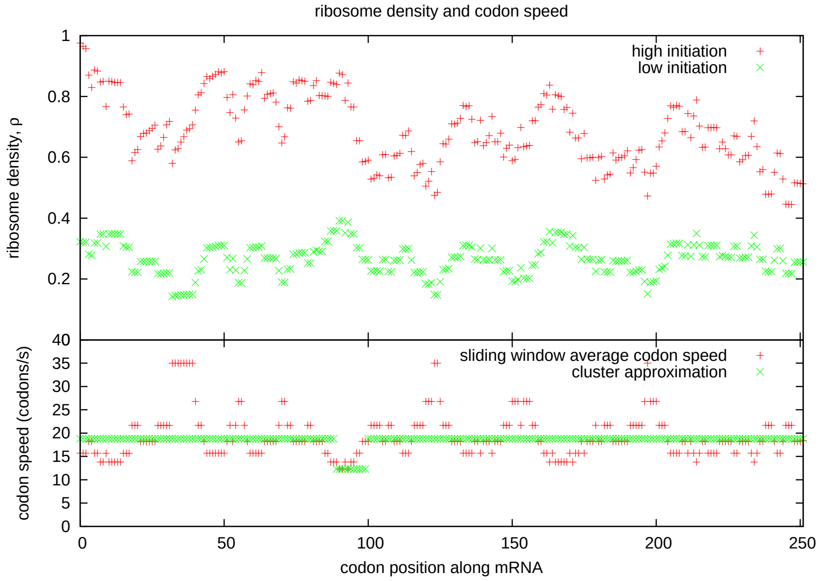


Figure 3.11: The bottom graph shows calculations average codon speed using a sliding window of size $l = 11$. The speed profile of the cluster approximation is also shown, the slowest site from the sliding window calculation is where the cluster of slow codon starts. The top graph shows ribosome density along the mRNA for high and low initiation rate.

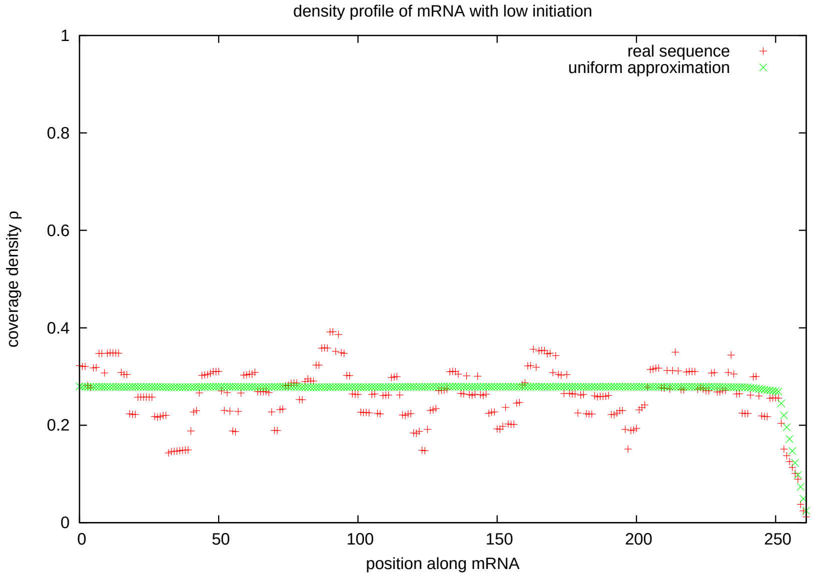


Figure 3.12: This is the density profile of the uniform TASEP approximation to the real seqence at low initiation rate ($A = 0.625$). Each point on the graph represents a gene.

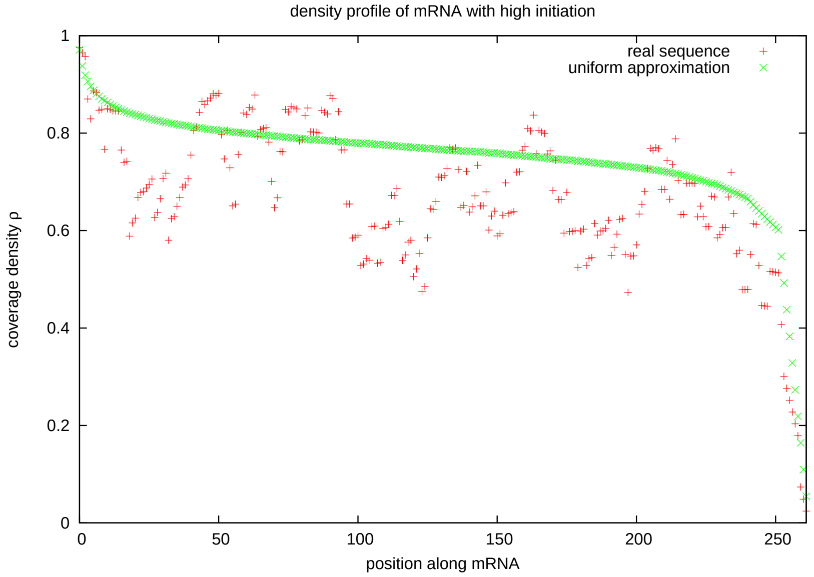


Figure 3.13: This is the density profile of the uniform TASEP approximation to the real seqence at high initiation rate ($A = 35.0$). Each point on the graph represents a gene.

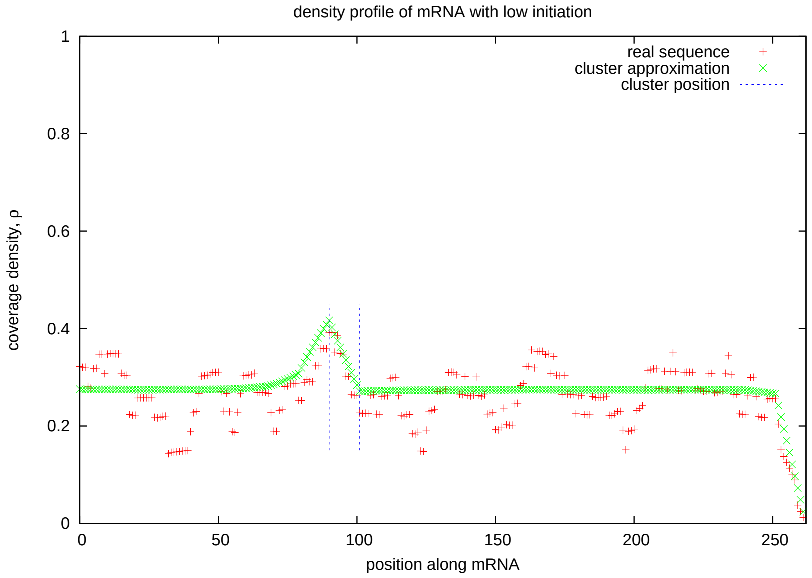


Figure 3.14: This is the density profile of the uniform TASEP approximation to the real seqence at low initiation rate ($A = 0.625$). Each point on the graph represents a gene.

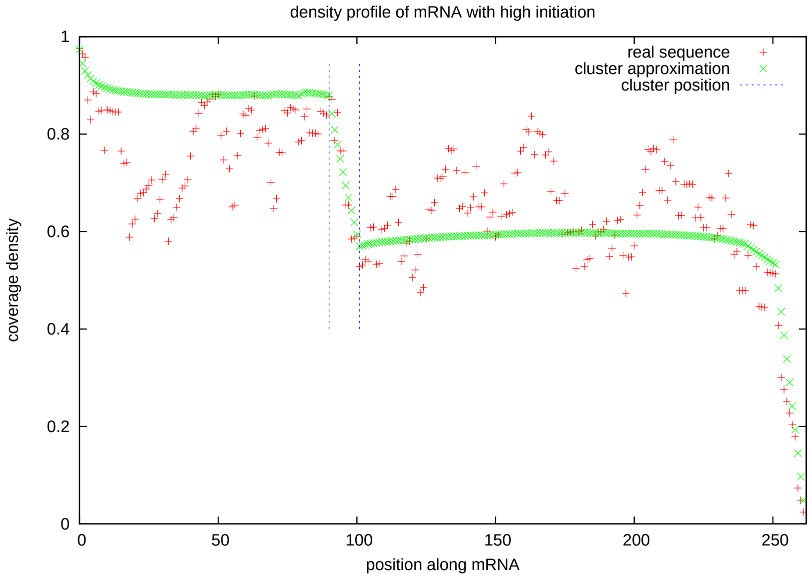


Figure 3.15: This is the density profile of the cluster TASEP approximation to the real seqence at high initiation rate ($A = 35.0$). A major drop around codon 145 is observed in both profiles. Each point on the graph represents a gene.

Ramps

The ramp is an observed lowering of adaptation values at the beginning of genes. The mechanism of 'ramping' is observed in Tuller et al (2010) and a similar signal was detected by us in *E. coli* as shown in figure 3.2 and 3.3, and is under investigation. The ramping mechanism is a region of slow codons which spaces out ribosomes entering the rest of the mRNA, preventing collisions, decrease idling time for ribosomes, and the probability of ribosomes falling off the transcript (Tuller et al, 2010). As seen in section 2.6, the current of ribosomes also decreases, which may not be desired. Figures 3.2 and 3.3, show a genome wide trend of a dip in codon adaptation and codon rates at the beginning of genes. This dip in codon adaptation was observed by Bulmer (1988) and in tRNA selection by Tuller et al (2010). The three curves correspond to averages over three ranges of δ_{genes} that have high, intermediate, and low adapted codon usage. The high expression genes have a higher average value of *delta* and codon speed, and the dip at the beginning is also most apparent.

The length of the ramp was estimated to be 1-3 ribosomes when tAI profiles were averaged over all genes in a large set of prokaryotic, and eukaryotic genomes (Tuller et al. 2010). Our data for approximating the ramp in real genes is at the moment incomplete. We have however looked at the position of where the slowest cluster occurs on the genes of the *E. coli* genome. We found that in high expression genes $\delta > 0$, 27.6% has the start of their slowest cluster at the first $l = 11$ codons, genes with $-0.3 < \delta \leq 0$, 15.9%, and $\delta \leq -0.3$, 12.9%. The counting error to those genes with $\delta > 0$ is greater than the low expression genes since only 228 genes in the high expression were sample, compared that in the low expression (1524 genes for $-0.3 < \delta \leq 0$, 2510 genes for $\delta \leq -0.3$). It is possible that a ramp and a cluster may both be present, but the slowest codons determine translation behaviour for sufficiently high kinetics of initiation. The ramp should also be sufficiently slow to be effective, so that a cluster of slow does not cause a transition. In order for the ramp to control translation, $\frac{v_{ramp}}{v_{ave}}$ must sufficiently deviate from 1. We see that for the genes chosen (see table A.5) that a ramp region of length 11 shows a weak $\frac{v_{ramp}}{v_{ave}}$ signal. In section 3.5.2, we discussed ribosome loading frequency of the lacZ gene of *Escherichia coli* from experiments, and we estimated the gene to be translated with a low initiation rate. When the initiation rate is low however, the ramp causes an even lower initiation rate, and does not necessarily improve translational efficiency.

Chapter 4

Multiple TASEPs and Finite Resource

In a cell, translational resources are shared between mRNA transcripts. The efficiency of translation should also take into account of the translational apparati (Kurland, 1991), such as the number of ribosomes. Two competing processes are at play when there are finite resources available for translation: the kinetics of initiation, and the rate of elongation. Ribosomes that are bound to mRNA can only translate that mRNA, and is excluded from initiating on other mRNAs. Slow elongation rates increase the time that ribosomes spend on transcripts, leading to fewer free ribosomes. The scarcity of ribosome in the cytoplasm leads to less opportunity for mRNA binding, so that initiation is lowered. As seen in the chapters before, when initiation is low, it limits translation and the number of ribosomes on transcripts. Bulmer (1991) concluded that the rate of protein production can change with the usage of codons only when the rate of elongation can change the kinetics of intiation. The TASEP model is extended for a finite number of ribosomes to test these assertions.

4.1 Multiple TASEPs

Multiple mRNAs can be simulated using the Gillespie algorithm. When multiple lattices are simulated, a separate rate is assigned to ribosome initiation on each mRNA. In addition, a finite pool of ribosomes may be used, and that mRNAs may have different codon configurations leading to competition for ribosomes. The possible events here are,

i) A ribosome hops.

ii) A ribosome enters a lattice.

iii) A ribosome leaves a lattice.

At each Gillespie simulation step, only a single event is chosen. The probability (p_j) of an event happening depends the kinetic rate for that event (r_j), and the total rate for any event happening ($R = \sum_j r_j$), and is the ratio $p_j = \frac{r_j}{R}$. A list of all ribosomes is generated in the order of lattice then position on the lattice. Similar to the single TASEP simulation, each event is checked if it could happen; which ribosomes are not blocked and can hop, and which mRNAs have the first l sites free and can bind a new ribosome.

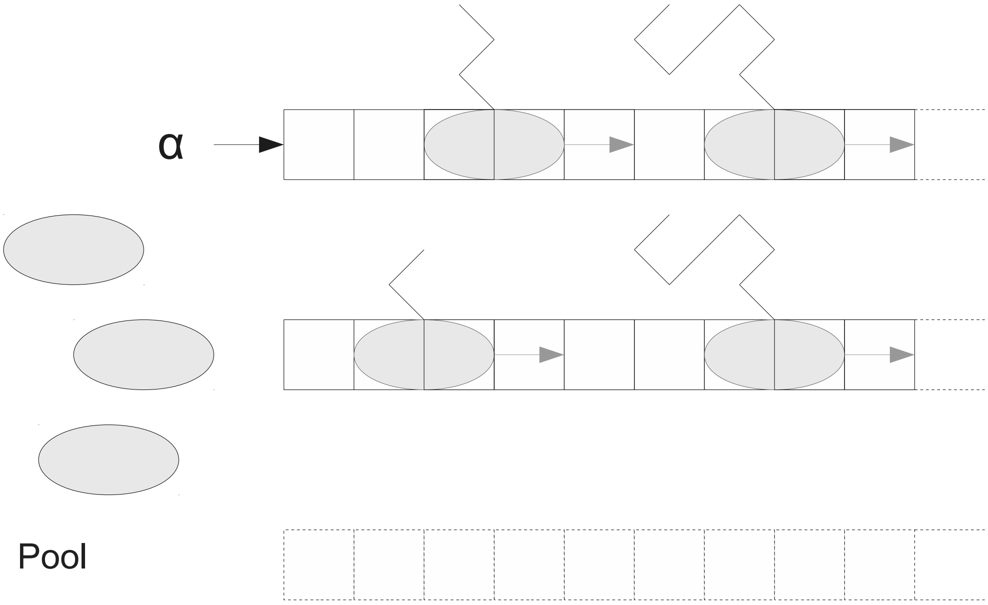


Figure 4.1: Multiple lattices can be simulated. Only one event, either initiation, termination or hopping may occur for the collection of lattices. A finite pool of ribosomes is also depicted.

4.2 mRNA Translation with Finite Resource

Let us define the following quantities in a cell,

N_{tot} , the total number of ribosomes in a cell;

N_f , number of free ribosomes in the cytoplasm;

N_{mRNA} , number of mRNA chains in a cell;

$\bar{n}$, ribosome number per mRNA.

So that at any given moment in time the cell has,

$$N_{tot} = N_f + N_{mRNA}\bar{n}, \tag{4.1}$$

number of ribosomes translating, or idle.

We propose to couple the collection of mRNAs through the TASEP initiation rate α on each chain, by the following relationship,

$$\alpha = k\,\frac{N_f}{N_{tot}}.$$

k is a kinetic parameter determined by the chemistry of the cell, and is the maximum binding rate for ribosomes. α here is the instantaneous rate of initiation at a given moment. Similar to chemical kinetic models, the rate of ribosome initiation/binding on an mRNA increases linearly with the fraction of free ribosomes. This describes increased opportunities for ribosome initiation in the cytoplasm. We also see that when the cytoplasm is sparse of ribosome the initiation rate approaches 0, as expected.

In the steady state we should be able to solve for the bound fraction corresponding to a given k. The pool of unbound ribosomes would have settled at $\bar{N}_f$ so that we can write the average initiation rate during steady state/at long times to be $\langle\alpha\rangle_{SS} \equiv \alpha = \frac{k}{N_{tot}}\bar{N}_f$. Rewriting equation 4.1,

$$N_{mRNA}\bar{n} = N_{tot}\left(1 - \frac{\alpha}{k}\right) \tag{4.2}$$

$$\bar{n} = n^*\left(1 - \frac{\alpha}{k}\right) \tag{4.3}$$

$$\tag{4.4}$$

where $n^* \equiv \frac{N_{tot}}{N_{mRNA}}$. The solution is shown schematically in figure 4.2.

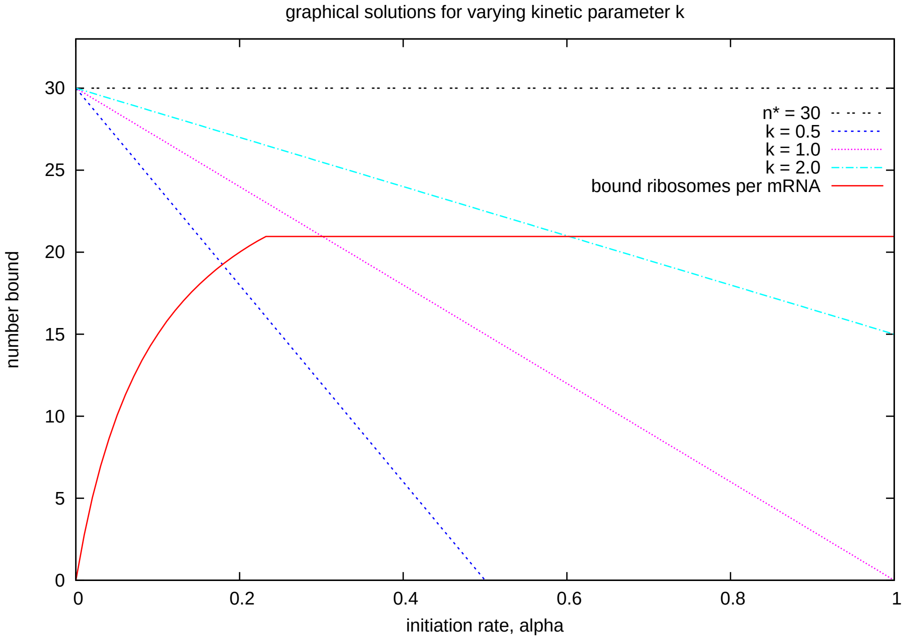


Figure 4.2: Graphical solution for linear initiation rate. The maximum number of bound ribosomes is $n^* = 30$. The solution to equation 4.2 is where a straight line intersect with the curve for ribosome number.

This solution represents the steady state that is reached by the two competing process of initiation and elongation. That the initiation rate is determined by the elongation rate, and vice versa. The number of free ribosomes is stable, leading to a well-defined initiation rate. The graphical solution interpolates the single TASEP model, and the theory incorporates finite resources with multiple TASEPs. We have compared simulation, the interpolation method, and the theory as shown in 4.3. We also note that n^* is the maximum possible bound ribosomes per mRNA.

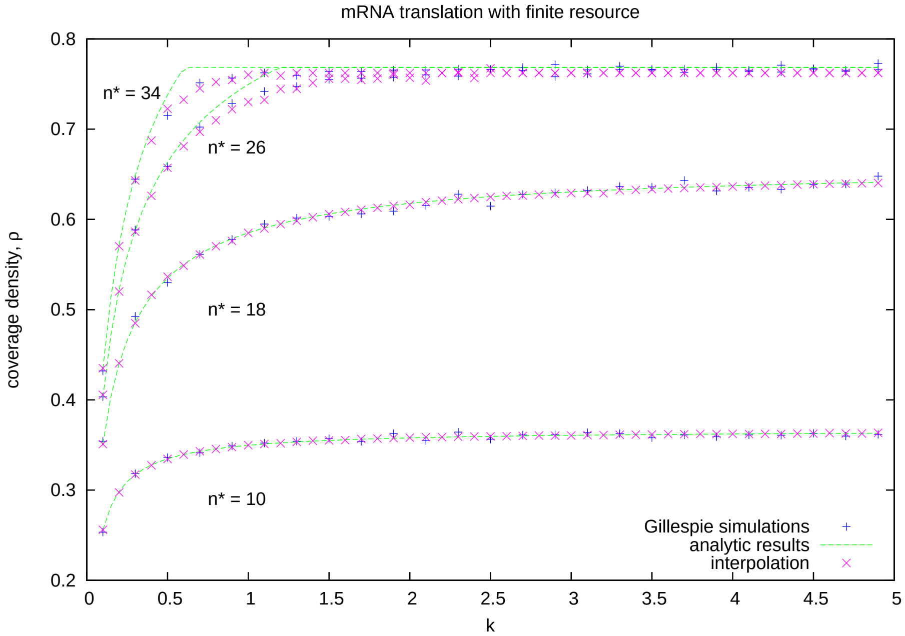


Figure 4.3: Translation of a collection of identical uniform rate mRNAs. The simulation is between 5 identical uniform TASEP, the interpolation is an application of graphical solution method using recursion, and the theoretical solution with linear initiation for a finite pool of ribosomes are shown. The average coverage density is shown for $n^* = 10$, 18, 26, 34. The higher curves correspond to higher n^*.

Figure 4.3 shows the mRNAs to enter the high density phase, where the ribosome density becomes independent of the initiation kinetics k only for when $n^* = \frac{N_{tot}}{N_{mRNA}}$ is high. When the total number of ribosomes in the system is low, mRNAs do not enter the high density phase, even for high initiation kinetics k. In this low density phase, the mRNAs are limited by the number of ribosomes in the cell. This can be a cause for low initiation rate found in a cell, where the translational resources shared between mRNAs are low. On the other hand, a low kinetic rate can also cause ribosomes to be used inefficiently, and many are required to maintain a fixed initiation rate. According to various estimates, there are as many as 20 000 ribosomes for 2000 mRNAs ($n^* = 10$)) in *E. coli* (Phillips et al, 2008), and similar numbers for *S. cerevisiae* with 200 000 ribosomes and 15 000 mRNAs (Warner, 1999).

Chapter 5

Conclusions and future investigations

5.1 Conclusions

In this thesis, we have tried to answer the question of how variation in codon adaptation affects the current J or the rate of protein completion. In our investigation we saw that the average codon rate v_{ave} of an mRNA increases with the usage of those codons preferred in the high expression genes. This is significant since, the average codon rate of highly expressed genes can be almost 3 times faster than low expression genes. At low initiation rate, where ribosomes do not interact, a higher mean codon rate ensures efficient translation.

The rate of initiation is an important translational control in that it can cause genes to go into a queueing phase which limits the rate of translation. The initiation rate required for transition to the queueing phase is found to be higher for high expression genes. A slow cluster leads to queueing at high initiation rate $A > A^*$, and the ribosome current becomes limited by the cluster.

We found that when the initiation rate is high, real genes are well approximated by a slow cluster and that the slowest cluster is an important feature of the mRNA. A cluster of slow codons has little to no effect on translation until a critical initiation rate A^* is reached, and the queueing phase begins. The critical initiation rate is different for different genes, and we found that high expression genes are more robust against queueing. High expression genes have higher critical initiation rate than low expression ones. In this sense, queueing in high expression genes also appears to be selected against.

At low initiation, genes do not queue, and the current of ribosomes completing translation of an mRNA is not limited by a slow cluster. Results from Kennell and Riezman (1977), Sorensen, and Pedersen (1991), and Mitarai et al (2008), were consistent with translation in the low density regime. That is, ribosomes do not form queues on mRNA, and the dynamics are well approximated by a uniform TASEP model. In this case, we conclude that high expression genes are unlikely to queue in real cells, and the ramp in high expression genes are not likely the result of selection against queueing.

This work has taught us that the slowest sequence of codons on an mRNA can cause queueing. The detailed study of queueing behaviour using the TASEP model tells us that queueing occurs only after a critical initiation rate is reached. We saw that well adapted genes tend to have a high critical initiation rate for queueing, and can accomodate more translating ribosomes before a queue would form. We find that in the kinetic picture, high expression genes have high average codon speeds, and are also less likely to queue. The position of the slowest codons is important, and we see that in the case of a ramp, where slow codons are at the beginning, the translation behaviour is different from a cluster in the bulk. The ramp spaces out ribosomes and prevent queues (Tuller et al. 2010) only when the initiation rate of ribosomes is high. The complicated interplay between codon adaptation, initiation kinetics, and ribosome interaction is successfully modeled by TASEP. The kinetic model can guide the way we think about how codon usage can affect the expression of proteins (codon configuration, average codon usage on an mRNA), and the evolutionary selection for translational efficiency (initiation and queueing).

5.2 Future investigations

Gene expression can be better approximated when there are multiple mRNAs translating a finite number of ribosomes. Using a simulation with multiple TASEP will allow us to simulate the effects of codon selection on different genes in a kinetic environment similar to a cell. In multiple TASEP, the initiation rate is limited by the number of ribosomes, and the chemical kinetics. The number of ribosomes on a single mRNA will be determined by the number of free ribosomes in the pool. When ribosomes are scarce and mRNAs have different codon adaptation, competition for finite resources emerges. Different genes have different codon usage, and the cell seek to optimize translation with mRNAs that have different codon adaptation.

More work is needed to be done on the initiation and ramp interaction. This work suggests that codon adaptation at the beginning of genes may be caused by other effects, such as RNA structure. Slow codons may be used at the beginning of genes to prevent secondary structures since the presence of secondary structures can inhibit initiation, or impede ribosome motion and cause codons to be translated slowly. Structure folding can be simulated on a TASEP lattice where the initiation rate or codon rates can change during simulation according to whether an unfolding event has occurred. We can estimate this rate based on the free-energy of RNA structures. The structural free-energy can be reliably calculated by existing programs.

Appendix A

Codon table, and list of genes

codon	amino acid	ϕ_H	ϕ_0	w	δ	anticodon	tRNA#	$W_{max}\times$tAI	Mitarai	thesis
UUU	Phe	0.228	0.574	0.295	-0.922	AAA	0	0.878	B	B
UUC	Phe	0.772	0.426	1	0.594	GAA	2	2	A	A
UUA	Leu	0.024	0.131	0.029	-1.688	UAA	1	1	B	B
UUG	Leu	0.04	0.128	0.048	-1.158	CAA	1	1.32	B	B
CUU	Leu	0.048	0.104	0.057	-0.764	AAG	0	0.439	B	B
CUC	Leu	0.038	0.104	0.045	-1.004	GAG	1	1	B	B
CUA	Leu	0.002	0.037	0.002	-2.901	UAG	1	1	B	C
CUG	Leu	0.847	0.497	1	0.534	CAG	4	4.32	A	A
AUU	Ile	0.249	0.508	0.332	-0.713	AAU	0	1.317	A	B
AUC	Ile	0.749	0.42	1	0.578	GAU	3	3	A	A
AUA	Ile	0.002	0.072	0.003	-3.446	UAU	0	1.306	C	C
AUG	Met	1	1	1	0	CAU	8	8	A	A
GUU	Val	0.52	0.259	1	0.697	AAC	0	0.878	A	A
GUC	Val	0.078	0.216	0.15	-1.02	GAC	2	2	B	B
GUA	Val	0.278	0.154	0.535	0.593	UAC	5	5	A	A
GUG	Val	0.124	0.371	0.238	-1.098	CAC	0	1.6	A	B
UCU	Ser	0.408	0.145	1	1.031	AGA	0	0.878	A	A
UCC	Ser	0.257	0.149	0.631	0.546	GGA	2	2	A	A
UCA	Ser	0.031	0.123	0.077	-1.37	UGA	1	1	B	B
UCG	Ser	0.013	0.154	0.031	-2.508	CGA	1	1.32	B	B

Table A.1: Codons, their frequency in a high expression set ϕ_H, their frequency in the genome ϕ_0, the ratio between the frequencies w, the log of the ratio $\delta = \ln(w)$, the anticodon isoacceptor corresponding to the codon, the number of tRNAs with the anticodon isoacceptor, the tRNA adaptation index for the codon, rate category according to Mitarai et al, 2008, and the category used in this thesis. W_{max} for *E. coli* is 8 for 8 methionine tRNAs.

codon	amino acid	ϕ_H	ϕ_0	w	δ	anticodon	tRNA#	W_{max}×tAI	Mitarai	thesis
CCU	Pro	0.142	0.158	0.197	-0.111	AGG	0	0.439	C	C
CCC	Pro	0.011	0.124	0.016	-2.376	GGG	1	1	C	C
CCA	Pro	0.126	0.191	0.176	-0.413	UGG	1	1	C	C
CCG	Pro	0.72	0.527	1	0.313	CGG	1	1.32	B	B
ACU	Thr	0.468	0.166	1	1.037	AGU	0	0.878	A	A
ACC	Thr	0.442	0.435	0.945	0.016	GGU	2	2	A	A
ACA	Thr	0.049	0.131	0.105	-0.981	UGU	1	1	B	B
ACG	Thr	0.041	0.268	0.088	-1.87	CGU	2	2.32	B	B
GCU	Ala	0.462	0.161	1	1.053	AGC	0	0.878	A	A
GCC	Ala	0.082	0.27	0.177	-1.196	GGC	2	2	B	B
GCA	Ala	0.269	0.214	0.582	0.229	UGC	3	3	A	A
GCG	Ala	0.188	0.356	0.408	-0.637	CGC	0	0.96	B	B
UAU	Tyr	0.238	0.569	0.312	-0.873	AUA	0	1.317	B	B
UAC	Tyr	0.762	0.431	1	0.571	GUA	3	3	A	A
UAA		0	0	0	0	UUA	0	1.306	B	B
UAG		0	0	0	0	CUA	0	1.306	B	B
CAU	His	0.299	0.572	0.426	-0.649	AUG	0	0.439	B	B
CAC	His	0.701	0.428	1	0.493	GUG	1	1	A	A
CAA	Gln	0.197	0.347	0.245	-0.568	UUG	2	2	B	B
CAG	Gln	0.803	0.653	1	0.207	CUG	2	2.64	A	A
AAU	Asn	0.123	0.451	0.14	-1.303	AUU	0	1.756	B	B
AAC	Asn	0.877	0.549	1	0.469	GUU	4	4	A	A

Table A.2: Codons, their frequency in a high expression set ϕ_H, their frequency in the genome ϕ_0, the ratio between the frequencies w, the log of the ratio $\delta = \ln(w)$, the anticodon isoacceptor corresponding to the codon, the number of tRNAs with the anticodon isoacceptor, the tRNA adaptation index for the codon, rate category according to Mitarai et al, 2008, and the category used in this thesis. W_{max} for *E. coli* is 8 for 8 methionine tRNAs.

codon	amino acid	ϕ_H	ϕ_0	w	δ	anticodon	tRNA#	W_{max}×tAI	Mitarai	thesis
AAA	Lys	0.715	0.766	1	-0.069	UUU	6	6	A	A
AAG	Lys	0.285	0.234	0.398	0.197	CUU	0	1.92	B	B
GAU	Asp	0.359	0.627	0.56	-0.559	AUC	0	1.317	B	B
GAC	Asp	0.641	0.373	1	0.543	GUC	3	3	A	A
GAA	Glu	0.763	0.69	1	0.101	UUC	4	4	A	A
GAG	Glu	0.237	0.31	0.31	-0.27	CUC	0	1.28	B	B
UGU	Cys	0.316	0.445	0.462	-0.342	ACA	0	0.439	B	B
UGC	Cys	0.684	0.555	1	0.209	GCA	1	1	A	A
UGA		0	0	0	0	UCA	0	1	B	B
UGG	Trp	1	1	1	0	CCA	1	1.32	A	A
CGU	Arg	0.683	0.38	1	0.586	ACG	4	4	A	A
CGC	Arg	0.302	0.4	0.442	-0.28	GCG	0	2.88	A	A
CGA	Arg	0.003	0.064	0.005	-2.951	UCG	0	0.0004	C	C
CGG	Arg	0.005	0.098	0.007	-2.97	CCG	1	1	C	C
AGU	Ser	0.047	0.151	0.115	-1.169	ACU	0	0.439	B	B
AGC	Ser	0.245	0.277	0.6	-0.125	GCU	1	1	B	B
AGA	Arg	0.007	0.037	0.01	-1.711	UCU	1	1	C	C
AGG	Arg	0	0.021	0.002	-3.446	CCU	1	1.32	C	C
GGU	Gly	0.621	0.338	1	0.61	ACC	0	1.756	A	A
GGC	Gly	0.359	0.404	0.578	-0.117	GCC	4	4	A	A
GGA	Gly	0.006	0.108	0.01	-2.813	UCC	1	1	C	C
GGG	Gly	0.013	0.151	0.021	-2.454	CCC	1	1.32	B	B

Table A.3: Codons, their frequency in a high expression set ϕ_H, their frequency in the genome ϕ_0, the ratio between the frequencies w, the log of the ratio $\delta = \ln(w)$, the anticodon isoacceptor corresponding to the codon, the number of tRNAs with the anticodon isoacceptor, the tRNA adaptation index for the codon, rate category according to Mitarai et al, 2008, and the category used in this thesis. W_{max} for *E. coli* is 8 for 8 methionine tRNAs.

δ	v_{ave}	v_{ramp}	$v_{cluster}$	length	note
0.434	26.9	26.8	21.7	78	murein lipoprotein
0.414	28.6	24.0	18.2	121	L7/L12
0.363	25.8	26.8	15.7	331	glyceraldehyde-3-phosphate dehydrogenase A
0.358	25.3	26.8	15.7	432	enolase
0.342	25.7	21.7	18.2	130	S9
0.319	22.4	26.8	15.7	241	S2
0.312	22.0	15.7	15.7	149	L9
0.296	23.3	24.0	12.3	234	L1
0.277	19.9	18.2	13.8	201	L4
0.267	20.1	15.7	10.1	273	L2
0.245	22.5	19.8	18.2	85	L27
0.239	21.1	21.7	15.7	82	S16
0.212	21.2	16.9	12.3	470	pyruvate kinase I
0.197	19.7	18.2	13.8	117	L18
0.180	19.2	12.3	12.3	131	S6
0.116	18.7	26.8	11.1	167	S5
0.114	18.4	15.7	12.3	262	enoyl-[acyl-carrier-protein] reductase
0.075	17.3	11.1	10.1	590	aspartyl-tRNA synthetase
0.048	17.9	12.3	9.3	388	succinyl-CoA synthetase, beta subunit
0.010	16.3	15.7	11.1	549	glucosephosphate isomerase

Table A.4: Genes selected from the genome of Escherichia coli. Their δ measure, average codon speed, ramp rate, cluster rate, and function are listed. There are ordered in increasing δ.

δ	v_{ave}	v_{ramp}	$v_{cluster}$	length	note
-0.129	14.6	10.1	8.6	1486	glutamate synthase, large subunit
-0.214	14.5	12.3	8.6	228	defective ribonuclease PH
-0.231	13.4	11.7	8.6	396	23S rRNA m(5)C1962 methyltransferase
-0.267	12.9	11.1	8.0	312	pseudouridine 5'-phosphate glycosidase
-0.273	12.7	8.0	8.0	226	RNase III
-0.280	12.7	13.8	8.6	295	glycerol-3-phosphate transporter subunit
-0.305	12.7	14.7	8.0	270	membrane ATPase of the MinC-MinD-MinE system
-0.321	12.7	10.6	7.5	165	none
-0.333	11.9	12.3	8.0	171	lipoprotein
-0.341	12.2	13.0	7.0	352	molybdate transporter subunit
-0.347	11.7	11.1	6.2	310	inner membrane protein, Predicted acyltransferas
-0.364	12.5	12.3	9.3	71	predicted phage lysis protein
-0.368	11.9	11.1	7.5	346	L-glyceraldehyde 3-phosphate reductase
-0.375	12.3	13.8	7.0	149	predicted flavoprotein
-0.413	12.3	13.8	8.6	138	export chaperone for FlgK and FlgL
-0.416	11.7	13.8	7.5	445	D-serine permease
-0.428	11.6	11.7	6.6	556	Hsp70 family chaperone Hsc62
-0.471	11.0	10.6	7.5	631	predicted inner membrane protein
-0.484	11.7	12.3	8.6	129	degenerate cysteine methyltransferase homolog
-0.556	10.7	9.3	6.2	121	inner membrane protein

Table A.5: Genes selected from the genome of Escherichia coli. Their δ measure, average codon speed, ramp rate, cluster rate, and function are listed. These are ordered in increasing δ.

Bibliography

Akashi H. (2003). Translational Selection and Yeast Proteome Evolution. Genetics, 164, 1291-1303.

Arava Y, Boas FE, Brown PO, Herschlag D. (2005). Dissecting Eukaryotic Translation and its Control by Ribosome Density Mapping. NAR, 33, 2421-2432.

Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D. (2003). Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. PNAS, 100, 3889-3894.

Bulmer M. (1988). 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.

Curran JF, Yarus M. (1989). Rates of Aminoacyl-tRNA Selection at 29 Sense Codons *in Vivo*. J. Mol. Biol., 209, 65-77.

Derrida B, Domany E, Mukamel D. (1992). An Exact Solution of a One-Dimensional Asymmetric Exclusion Model with Open Boundaries. J. Stat. Phys., 69, 667-687.

de Smit MH, van Duin J. (1990). Secondary structure of the ribosome binding site determines translational efficiency: A quantitative analysis. Proc. Natl. Acad. Sci. USA, 87, 7668-7672.

dos Reis M, Wenisch L, Savva R. (2003). Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. Nucleic Acids Research, 31, 6976-85.

dos Reis M, Savva R, Wernisch L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Research, 32, 5036-5044.

Gillespie DT. (1977). Exact Stochastic Simulation of Coupled Chemical Reactions. J. Phys. Chem., 81, 2340-2361.

Gu W, Zhou T, Wilke CO. (2010). A Universal Trend of Reduced mRNA Stability near the Translation-Initiation Site in Prokaryotes and Eukaryotes. PLoS Comput. Biol., 6, e1000664.

Higgs PG, Ran W. (2008). Coevolution of Codon Usage and tRNA Genes Leads to Alternative Stable States of Biased Codon Usage. Mol. Biol. Evol., 25(11), 2279-2291.

Higgs PG, Ran W. (2012, under review).

Ikemura T. (1981). Correlation between the Abundance of *Escherichia coli* Transfer RNAs and the occurrence of the Respective Codons in its Protein Genes. J. Mol. Biol., 146, 1-21.

Kang C, Cantor CR. Structure of Ribosome-bound Messenger RNA as Revealed by Enzymatic Accessibility Studies. J. Mol. Biol., 181 241-251.

Kennell D, Riezman H. (1977). Transcription and Translation Initiation Frequencies of the *Escherichia coli* lac Operon. J. Mol. Biol., 114, 1-21.

Kolomeisky AB. (2001). Exact results for parallel-chain kinetic models of biological transport. J. Chem. Phys., 115, 7253-7259.

Kudla G, Murray AW, Tollervey D, Plotkin JB. (2009). Coding-Sequence Determinants of Gene Expression in Escherichia coli. Science, 324, 255-257.

Kurland CG. (1991). Codon bias and gene expression. Federation of European Biochemical Societies, 285, 165-169.

Lakatos G, Chou T. (2003). Totally asymmetric exclusion processes with particles of arbitrary size. J. Phys. A: Math. Gen., 36, 2027-2041.

MacDonald CT, Gibbs JH, Pipkin AC. (1968). Kinetics of Biopolymerization on Nucleic Acid Templates. Biopolymers, 6, 1-25.

MacDonald CT, Gibbs JH. (1969). Concerning the Kinetics of Polypeptide Synthesis on Polyribosomes. Biopolymers, 7, 707-725.

Mitarai N, Sneppen K, Pedersen S. (2008). Ribosome Collisions and Translation Efficiency: Optimization by Codon Usage and mRNA Destabilization. J. Mol. Biol., 382, 236-245.

Nagel K, Schreckenberg M. (1992). A cellular automaton model for freeway traffic. J. Phys. I France, 2, 221-2229.

Ninio J. (2006). Multiple stages in codon-anticodon recognition: double-trigger mechanisms and geometric constraints. Biochimie, 88, 963-992.

Percudani R, Pavesi A, Ottonello S. (1997). Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. J. Mol. Biol., 338, 439-444.

Phillips R, Kondev J, Theriot J. (2008). Physical Biology of the Cell. Garland Science.

Duret L. 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. Trends in Genetics, 16, 287-289.

Ramakrishnan V. (2002). Ribosome Structure and the Mechanism of Translation. Cell, 108, 557-572.

Ran W, Higgs PG. (2010). The Influence of Anticodon-Codon Interactions and Modified Bases on Codon Usage Bias in Bacteria. Mol. Biol. Evol., 27, 2129-2140.

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

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. (2005). Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Research, 33(4), 1141-1153.

Shaw LB, Kolomeisky AB, Lee KH. (2004). Local inhomogeneity in asymmetric simple exclusion processes with extended objects. J. Phys. A: Math. Gen., 37, 2105-2113.

Shaw LB, Zia RKP, Lee KH. (2003). Totally asymmetric exclusion process with extended objects: A model for protein synthesis. Phys. Rev. E, 68, 021910.

Slayter H, Kiho Y, Hall CE, Rich A. (1968). An electron microscopic study of large bacterial polyribosomes. J. Cell Biol., 37(2), 583-590.

Sørensen MA, Kurland CG, Pedersen S. (1989). Codon Usage Determines Translation Rate in *Escherichia coli*. J. Mol. Biol. 207, 365-377.

Sørensen MA, Pedersen S. (1991). Absolute *in Vivo* Translation Rates of Individual Codons in *Escherichia coli:* The Two Glutamic Acid Codons GAA and GAG Are Translated with a Threefold Difference in Rate. J. Mol. Biol. 222, 265-280.

Steitz JA. (1969). Polypeptide Chain Initiation: Nucleotide Sequences of the Three Ribosomal Binding Sites in Bacteriophage R17 RNA. Nature, 224, 957-963.

Tripathi T, Schuetz GM, Chowdhury D. (2009). RNA polymerase motors: dwell time distribution, velocity and dynamical phases. J. Stat. Mech., 8, P08018.

Takanami M, Zubay G. (1964). An estimate of the size of the ribosomal site for messenger RNA binding. Proc. Natl. Acad. Sci. U.S., 51, 834-839.

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, 1-11.

Warner JR. (1999). The economics of ribosome biosynthesis in yeast. Trends in Biochemical Sciences. 24, 437-440.