ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE
FACULTE DES SCIENCES DE LA VIE

Projet de master en Bioingénierie et Biotechnologie

A Stochastic Translation Algorithm: The View Of The Ribosome

Réalisé par:
Olivier Burri

sous la direction du
PROF. VASSILY HATZIMANIKATIS
Au Laboratory of Computational Systems Biology (LCSB)
de l'EPFL

Expert Externe Prof. Babatunde A. Ogunnaike,

LAUSANNE, EPFL 2009

Contents

Abstract

This work deals with the creation of a stochastic translation algorithm capable of encompassing the reactions for translation initiation, elongation and termination in a unified framework based on Gillespie's algorithm. By looking at reactions from the point of view of the ribosome. That is as transitions from one of the 64 available codons to the next, the system was reduced to 64+2m equations, with m being the number of mRNA species in the system. Using this approach, the system no longer scales with molecule numbers or mRNA length, increasing only by 2 reactions for each additional mRNA species.

The algorithm was validated by replicating the results from the Heinrich-Rapoport (H-R) model as well as the Zouridis-Hatzimanikatis (Z-H) model. The stochastic protein translation model by Mitarai *et al.* was also recovered using the algorithm. Furthermore the use of the stochastic translation algorithm allows for a complete analysis of the noise of the system. In the H-R model it is shown that increased noise levels in polysome sizes around the phase shift correspond to the fast increase in polysome size observed in their results. From comparing the Z-H model to the stochastic translation algorithm, it was shown that as protein production reaches its maximum with respect to polysome size, protein noise reaches a minimum.

The report ends by suggesting further work that can be accomplished by using this algorithm and the questions that can be answered, opening the door to several interesting experiments.

Lausanne, June 19^{th} 2009,
Olivier Burri

Chapter 1

Introduction

Chapter 2

Goal Of This Project

The Laboratory of Computational Systems Biology has accumulated an extensive knowledge about protein translation through the use of deterministic models [10, 13, 20, 21]. To enhance this knowledge and complete the set of tools and models available, the need for a stochastic model for translation became apparent. There exsits a large body of literature regarding the study of noise in biological systems [1, 7, 9, 11, 12, 17], but for the purposes of the current approach, the models suggested were either too global or scaled directly with mRNA length [18]

The aim of this project was to develop a stochastic simulation algorithm (SSA) for prokaryotic protein translation that could encompass the variable translation rates of the ribosomes as they pass over different codons. The algorithm was to provide information that allowed for the study of noise distributions in the system, which would help to understand the engineering principles driving translational regulation.

The approach focused on a Gillespie direct method algorithm, where the elongation reactions are seen from the point of view of the ribosomes, switching from one codon x to the next y on the mRNA as $R_x \rightarrow R_y$ with elongation rate $\mathrm{K}_E(x)$. To select which ribosome reacts at each iteration of the elongation process, the index of a ribosome is selected at random from a table storing the position and mRNA molecule of each ribosome at a reacting codon x. This allows for the algorithm to be independent of the mRNA length, allowing several mRNA species to be included simultaneously if desired by adding only 2 reactions per new mRNA species to the Gillespie backbone.

To validate the algorithm, it was used to recover the results from the continuous H-R [6] and Z-H [20] models, and from the Monte Carlo model

of Mitarai et al [15].

The algorithm is able to produce information about every species in the system, the ribosomal distributions along the length of the mRNA, the number of initiation, elongation, termination reactions, the polysome sizes. This data is output to text files that can then be easily read by a statistics software suite.

2.1 On Systems Biology

"You Can Build A Perfect Machine Out Of Imperfect Parts."

This was a"flavor text" that had been printed on a card from a role playing game in the late 90's. It is a nicer take on "more than the sum of it's parts" and goes quite well with systems biology.
Biological systems are inherently complex. The amount of data that exists inside a single 1.10^{-15}l cell is unimaginable, The entirety of the organism's genetic code is stored digitally within it, as well as an extremely dense proteomic machinery that can read this code, respond to external stimulus, maintain homeostasis, coordinate sequences of intra- and extra- cellular reactions, eliminate harmful substances and micro-organisms, as well as perform a myriad of other actions; this make the cell one of the most beautiful and complex systems that are known to exist. And yet it is made of of common atoms, forming molecules. These molecules are subjected to thermodynamic and entropy laws that break them down over time. But put together these "imperfect parts" that do not amount to much on their own, contribute to the most amazing emergent behavior that is known: Life.

Looking at each individual component through a reductionist approach, though yielding very important information, cannot explain the behavior of a whole system. Systems biology's paradigm adopts a holistic view of biology, which seeks to study the complex interactions between molecules, enzymes, DNA, proteins, etc... by describing it as a system. This approach can help derive the "engineering laws" that govern biological systems of many scales and explain their properties.

A beautiful example is a recent article by Grubelnik *et al.* describing the striking similarities between biological and electrical systems with the use of signal amplification cascades [4]. The paper illustrates how the design principles for signal amplification cascades in cells are mathematically fully

equivalent to those of man-derivied amplifier systems.
With this idea in mind, discovering new principles through modeling can help us gain a deeper understanding of the system in question, but also derive new laws that can be applied in a number of engineering fields.

How do deterministic behaviors, such as delicate gene regulation, chemostasis or signal transduction .arise from such an inherently noisy system? One needs to remember that this is a physical system, ruled by thermodynamics, and that the small number of molecules involved make the noise a non-negligible part of it.

Systems biology has managed to strive thanks to the development of the computer to aid in the calculations of the models that would be, except in the simplest cases, almost impossible to work out analytically.

2.2 Protein Translation in Prokaryotes

Prokaryotic protein translation is quite well understood in terms of the players involved and an important body of literature exists, providing experimental results that can help create mathematical models.
Here a brief overview of prokaryotic translation is presented, so as to understand the system that is to be modeled by this paper.[1]. The initiation of translation is summarized in figure2.1. The two main steps are the high affinity bindings of the ribosomal 30s subunit to the 5' end of a free mRNA strand, followed by the binding of the ribosomal 50s subunit along with the tRNA carrying methionine, placed on the ribosomal A site.

Ribosomes are large ribosomal RNA/ ribosomal protein (65%, 35%) complexes, of a length of about 20nm, weighing about 2700kD[2]. Due to their size, the ribosomes occlude a certain number of codons on the mRNA when they are bound to it. This is referred to as their Occlusion Distance.

Because prokaryotes do not possess a nuclear membrane, DNA is translated into mRNA that quickly becomes bound to ribosomes as the mRNA is churned out from the RNA polymerase. As long as a given length of the 5' end of the mRNA is accessible (not occluded by a ribosome), another ribosome can bind to the same strand, resulting in a 1D row of ribosomes moving

[1]Wikipedia: http://en.wikipedia.org/wiki/Prokaryotic_translation
[2]http://redpoll.pharmacy.ualberta.ca/CCDB/cgi-bin/STAT_NEW.cgi

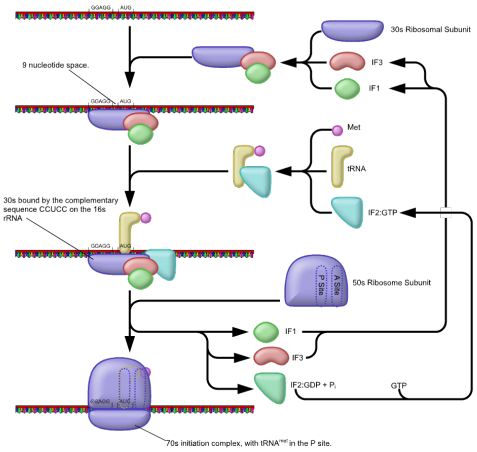

Figure 2.1: The initiation of protein synthesis in prokaryotes, in a condensed form. Image ©Richard Wheeler (Zephyris) 2005, from the free licensed media file repository Wikimedia.

along the mRNA, which is commonly known as a Polysome.

The elongation process involves several steps that are summarized by the model of Zouridis & Hatzimanikatis [20, 21] (See Figure 2.3). For the purpose of this work, it is enough to notice that the elongation process revolves around the 1D movement of ribosomes along the length of the mRNA, which each step matching a codon to its anti-codon on the tRNA, and adding an amino acid to the growing protein chain. Each codon that makes up the natural genetic code of most organisms, consists of a triplet of any of the four nucleotides (A,U,C or G), yielding a total of 64 possible nucleotide combinations. Considering that there are 20 amino acids that are digitally coded this way, different codons can code for the same amino-acids. The apparent redundancy (or genetic code degeneracy) confers robustness to protein translation and also adds a new level of transcription regulation [3, 14, 16], through variable codon transcription rates for the same amino-acids, which will be discussed during this work. A table of the natural genetic code is available in the annex (Figure9.1).

The termination process occurs when a ribosome encounters a STOP codon while reading the mRNA. This causes the unbinding of the two ribosomal subunits and the liberation of the newly synthesized peptide chain.

2.3 Modeling Translation

Following is a very brief introduction to modeling in systems biology, in either deterministic or stochastic form.

Continuous Deterministic Models

For a long time, deterministic models have been the main focus of systems biology. The concept behind these is that a biological system or pathway consists of a series of inputs and outputs that can be expressed in terms of ingoing and outgoing fluxes for each species in the system. These fluxes represent the change in time of the different species in the system, and are expressed by using mass action kinetics, laws that allow to explain the behaviors of solutions in dynamic equillibrium. The flux expressions make up sets of ordinary differential equations (as well as other types) that can then be solved through numerical integration. They offer both quantitative and qualitative insights as to the behavior of the system over time and at steady state.

Example: Heinrich-Rappoport Model Heinrich and Rappoport [6] presented a deterministic model of protein translation that takes into account the physical consequence of ribosomes sitting on mRNA by attributing them an occlusion distance L. In their model, by varying the elongation rate K_E, the initiation rate K_I, and the termination rate K_T, they go on to show how ribosomes pile up on an mRNA strand, and how this affects protein production and polysome size. A sample result is shown on figure 2.2.

Example: Hermione In a paper by Hermione Zouridis and Vassily Hatzimanikatis [20, 21], a more in-depth deterministic, sequence-specific model is proposed, taking into account most of the known reactions of the elongation process explicitly. This allowed for the elongation rates to be codon-dependent as well as offering a more accurate model that made fewer assumptions, based on first principles rather than *ad-hoc*assumptions

A summary of the model is discussed in figure 2.3. The elegance in the approach lies in considering the ribosomes as molecules at different states (corresponding to different steps in the elongation process), and the development of control parameters to quantify the influence of each state on the system, gaining a better understanding of the limiting steps in the translation process, depending on the parameters used.

The conclusions drawn from their studies are as follows (non-exhaustive):

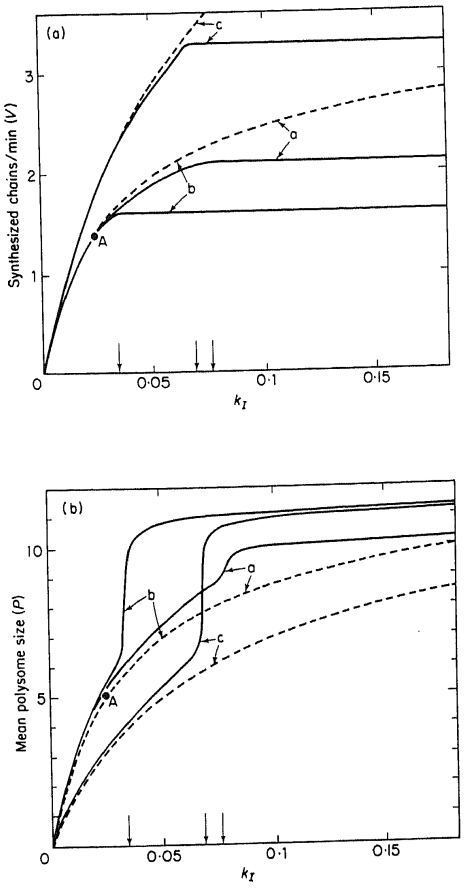

Figure 2.2: Synthesized proteins per minute and mean polysome sizes for α-globin with varying initiation rate (K_I)(x axis). Ignore the dashed lines. Values of the curves: (a) 1 mRNA, large ribosome excess (100), K_E=43.2min^{-1} , K_T = 6.0min^{-1}; (b) K_T = 6.0min^{-1}; (c) Kw = 90.0min^{-1}

- Increasing polysome sizes lead to increased protein translation rate, which reaches a maximum, and then decreases again, due to overcrowding.
- The kinetics are initiation and elongation limited for low polysome sizes and are termination limited for high polysome sizes.
- The maximum protein synthesis rate occurs at the polysome size corresponding to the maximum elongation rates for which polysome distributions are still uniform along the length of the mRNA.

These points are summarized in figure 2.4, taken from [20].

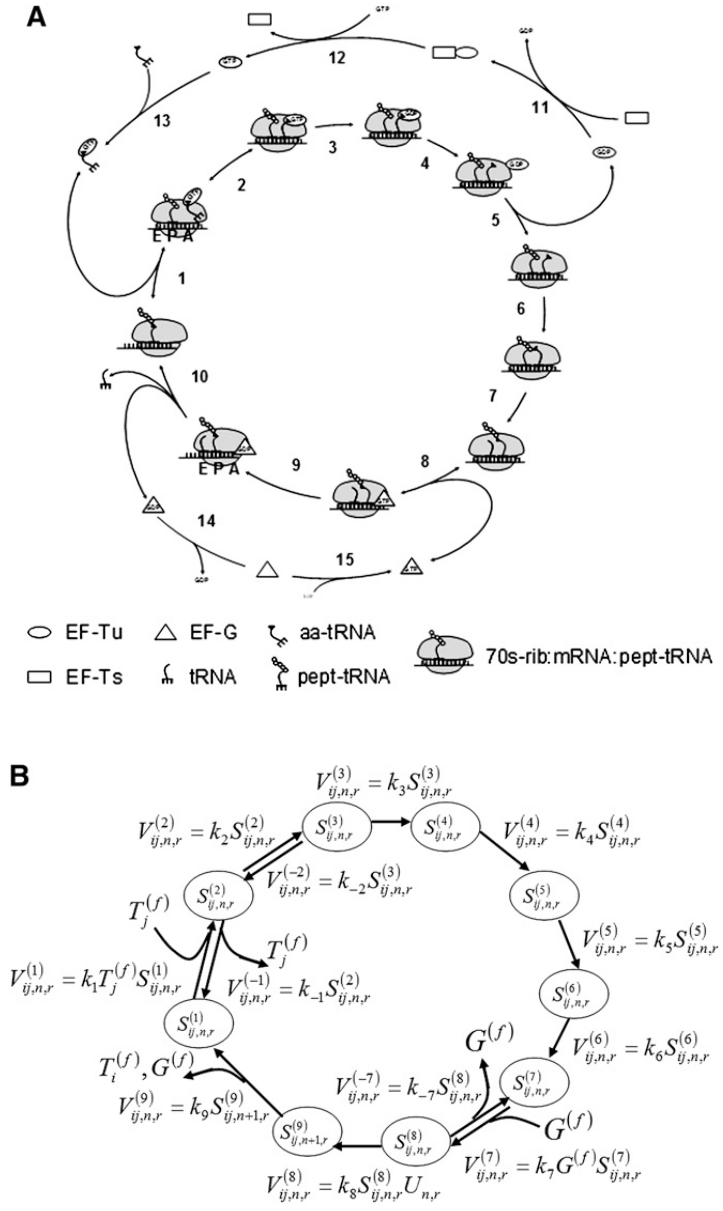

Figure 2.3: A: Schematic representation of the elementary steps of the elongation process. B: Graphical representation of the elongations steps considered in the Zouridis & Hatzimanikatis model.

While providing very valuable information on the description of the translational machinery, the Z-H model cannot capture the effects of noise since the processes are modeled as continuous. So information about the stochastic

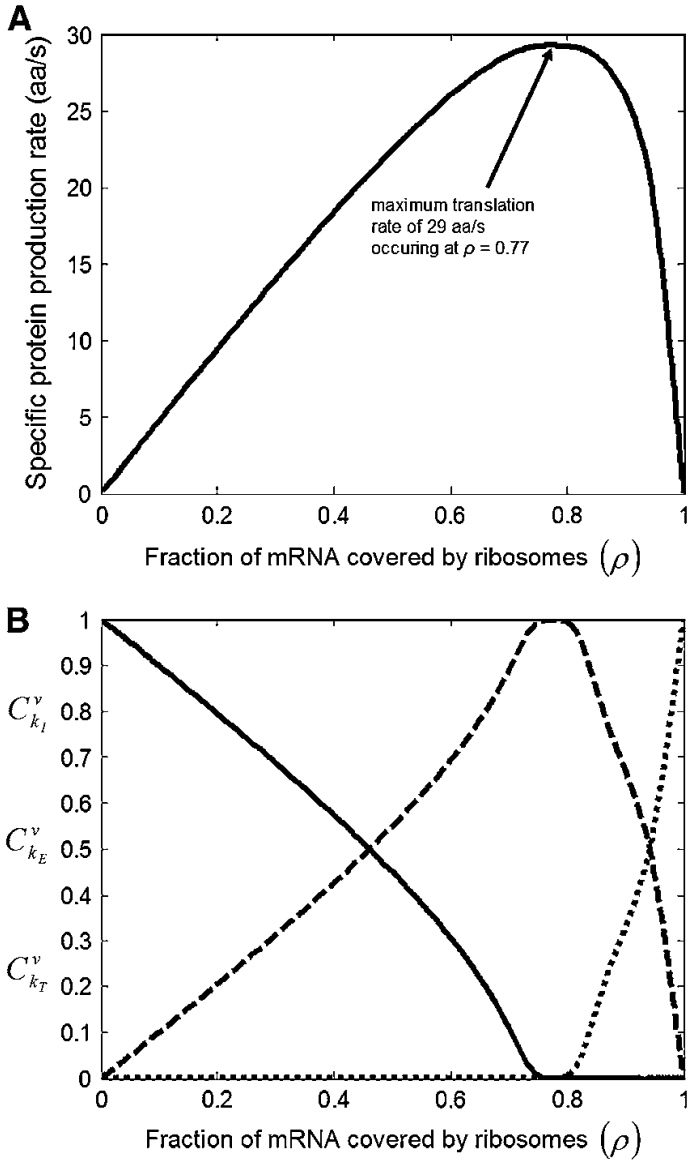

Figure 2.4: Protein production rates with respect to ribosomal density ρ. A: Plot showing maximum production rate. B: Control parameters for Initiation $C^v_{K_I}$, Elongation, $C^v_{K_E}$, and Termination $C^v_{K_T}$, and their relative influences as polysome sizes increase.

behavior of the system cannot be recovered.

Discrete Stochastic Models

Stopping for a moment to consider the nature of the systems under study, it can be shown that under certain conditions, modeling in continuous form is not sufficient [8]. Deterministic models are acceptable when faced with very large numbers of molecules, so that the way their fluxes are expressed approximate the molecular fluxes happening at the thermodynamic limit. In practice, a system is usually considered to be at the thermodynamic limit when its molecule numbers are close to the Avogadro number($6.022.10^{23}$ Molecules). Below this value, deterministic modeling is not sufficient because the discrete nature of the system generates stochastic behaviors.
Cellular processes are confined to a small volume (that of the cell) and to small molecule numbers, so many of them operate under the thermodynamical limit.

With about 4'000 mRNAs and 18'000 Ribosomes, a typical *E.coli* cell cannot be modeled in a deterministic way if emergent behavior caused by stochastic processes wants to be studied . In the case of protein translation, the discrete nature of the states of the molecules, such as the availabilities of free mRNAs for ribosome binding or the one-codon-at-a-time movement of the ribosome along an mRNA strand, require a discrete 'per step' view that reflects the finite nature of the system, also called a jump Markov Process.
To remedy this with the help of computers, methods for modeling discrete systems have been developed, and take many forms, such as Markov Models, Petri Nets, and Kinetic Monte Carlo

Monte Carlo methods rely on random number generators to simulate the non-deterministic nature of the systems being modeled. (For a good introduction, the book "Stochastic Modelling for Systems Biology" by Darren J. Wilkinson is suggested).

The difference between deterministic models and stochastic models, mathematically speaking, is that rather than considering the fluxes of molecules going from a given state or species to another, these models deal with the probabilities of these events taking place inside a certain volume with certain molecule numbers (not concentrations). One ends up with a linear system of differential probability equations, called the chemical master equation (2.1).

$$P\left(\vec{X};t+dt|\vec{X}_0;t_0\right)=P\left(\vec{X};t|\vec{X}_0;t_0\right)\left(1-\sum_{j=1}^{M}a_j dt\right)+\sum_{j=1}^{M}P\left(\vec{X}-\vec{v}_j;t|\vec{X}_0;t_0\right)dt \tag{2.1}$$

In this expression, $\vec{X}$ is the state vector of the system with $P\left(\vec{X};t|\vec{X}_0;t_0\right)$ the probability that the state is $\vec{X}$ at time t, given that the state was $\vec{X}_0$ at time t_0. The first expression in the right-hand side sum then represents the probability that no reaction takes place in the interval $t+dt$, and the second expression represents the probability of there being a reaction $\vec{v}$ so that the system was at $\vec{X}-\vec{v}_j$ at time t. M is the number of reactions in the system and $a_j = c_j.h_j$ is the propensity of reaction j, that is the probability that reaction j will occur within $t+dt$ in the volume considered. c_j is the mesoscopic reaction rate constant of reaction j and h_j is the number of possible combinations of the reacting molecules of reaction j.

One cannot solve the chemical master equation explicitly for anything but the smallest systems, as it grows combinatorially. What can be done using Monte Carlo methods is to recover a realization of the chemical master equation. By taking the data of many runs into $\vec{X}$ that follow the constraints of $P\left(\vec{X};t|\vec{X}_0;t_0\right)$, it is possible to infer the complete state-space of the system.

Example: Mitarai Paper This Masters project started by replicating the algorithm developed by Mitarai *et al.* [15], which uses a Monte Carlo simulation with a constant time step Δt. At each step, the probability of a ribosome binding to mRNA is calculated as $\mathrm{Ks}_f\Delta t$, the probability of bound ribosomes to move as $\mathrm{K}_E(\mathrm{codon})\Delta t$, the probability of a ribosome at position L to unbind from the mRNA as $\mathrm{K}_T\Delta t$. With Ks_f, $\mathrm{K}_E(\mathrm{codon})$ and K_T the initiation, codon-dependent elongation and termination rate constants. Figure 2.3 offers a schematic representation of the different steps involved in their model.
This paper also deals with a further aspect of protein translation which was mRNA rates of birth and decay, which will not be included here.

The model was tested against experimental data of recombinant protein production using the LacZ plasmid and a modified construct, where "slow" codons were added. The output they observed was protein production using radioactively labeled amino-acids.

By looking at the accumulated radioactivity in lacZ operon variants and using their Monte Carlo agorithm to replicate the radioactivity measure-

ments, they were able to show that the presence of ribosomal queueing due to variable elongation rates was the most likely factor explaining data (Figure 2.6).

Through their discussion, show that a fully deterministic or a uniform codon distribution for the stochastic mechanism were not sufficient to describe the model, explaining that the variable codon elongation rate constants, though creating bottlenecks close to slow codons, also helped regulate translation through careful codon usage.

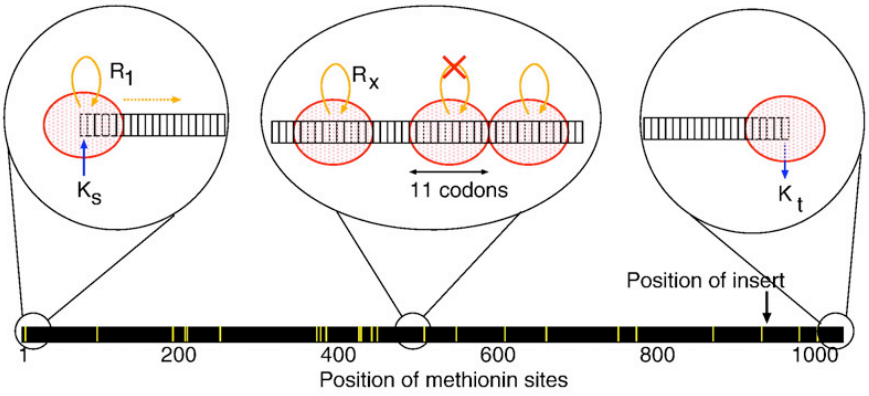

Figure 2.5: Schematic representation of Monte Carlo model that was used in the Mitarai paper. R_1, ..., R_x are the different codon-specific rates. Ks, the Initiation rate and Kt the termination rate.

Gillespie's Algorithms

The paper by Mitarai *et al.* is an elegant approach at modelling protein translation in a stochastic form, but has certain limitations that this work aims to overcome.

- Ribosomes are considered to be in excess, thus ribosomal concentrations play no role in the model. However, in *E.coli* , the concentration of ribosomes translates to about 18'000 per cell. If one considers 1000 mRNAs, that means that there are about 18 ribosomes available per mRNA on average. The similarity of scale demands for ribosomes to be considered explicitely.[3]

- The framework is a fixed time step Monte Carlo which does not capture the physical properties of the system, which consists of chemical

[3] *E.coli* Stats: http://redpoll.pharmacy.ualberta.ca/CCDB/cgi-bin/STAT_NEW.cgi

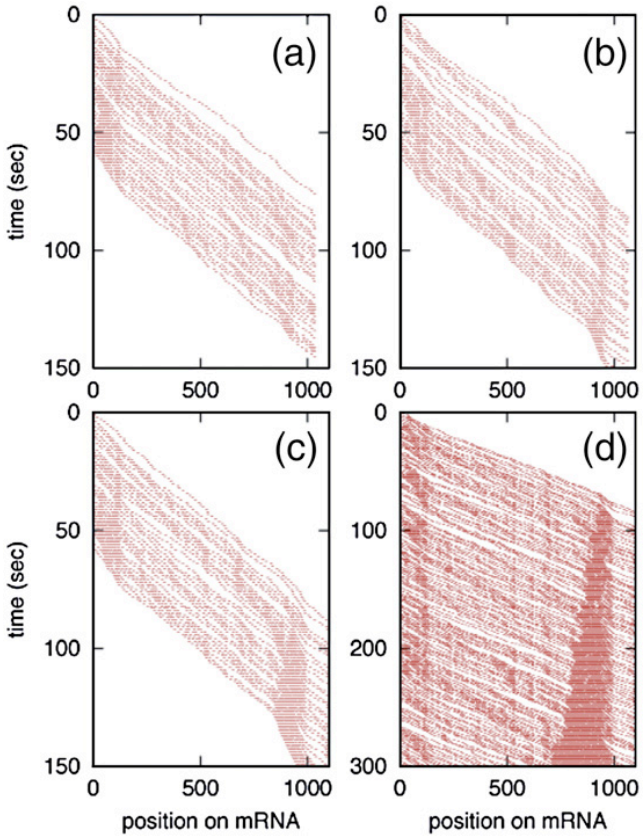

Figure 2.6: Space time plot of ribosome traffic for the LacZ Operon. For the wild type LacZ (a) and for the LacZ gene with an insert of slow codons at the end of the sequence. For figures a,b and c, we see how individual ribosomes (red lines) move along the mRNA over time for 30 proteins being translated. Figure d shows how a queue can develop on an mRNA.

reactions occurring due to molecular collisions in a given volume which carries with it some implications and should therefore have a framework that encompasses this physical system.

- Consider the 4'000 mRNAs in a typical *E.coli* cell. They are not all the same mRNAs and code for different genes at different rates. The developed algorithm should allow the study of mRNA competition for a shared pool of ribosomes among different species.

- Due to the microscopic nature of the system, noise is of importance. LiK$_E$ mentionned above, the study of noise propagation should be considered. Namely, how does initiation noise (Collisions of molecules in 3

dimensions, ribosomes binding to mRNA) relate to termination noise (unbinding of a ribosome after passing through a 1D system) and how would noise propagate in the system.

In 1977, Daniel T. Gillespie [2] proposed a stochastic simulation algorithm (SSA) that offered a framework in which the physical nature of the system was taken into consideration as well as an easy to implement method that served as the backbone for the rest of the algorithm to be written into. A quick summary of the algorithm is presented in figure 2.7.

How It Works (Direct Method) Gillespie's analysis shows that in a well-mixed discrete system driven by many non-reacting elastic collisions of molecules that bring reacting species together, the probability of two molecules colliding in an infinitesimal time can be rigorously calculated. He then proposed to break down the simulation into single reaction steps, where at each step the questions"When will the next reaction take place?" and "What kind of reaction will it be?" are asked. Using two random numbers, one exponentially distributed that gives the time to the next reaction, and the second uniformly disctributed that gives the type of reaction that will take place, Gillespie showed that it yielded an unbiased walk in the probability space of the system. This is called Gillespie's "Direct Method". Variants of this algorithm exist, such as the Next Reaction Method and First Reaction Method, and although they were not used in this work, they could be easily extended for the use of this algorithm.

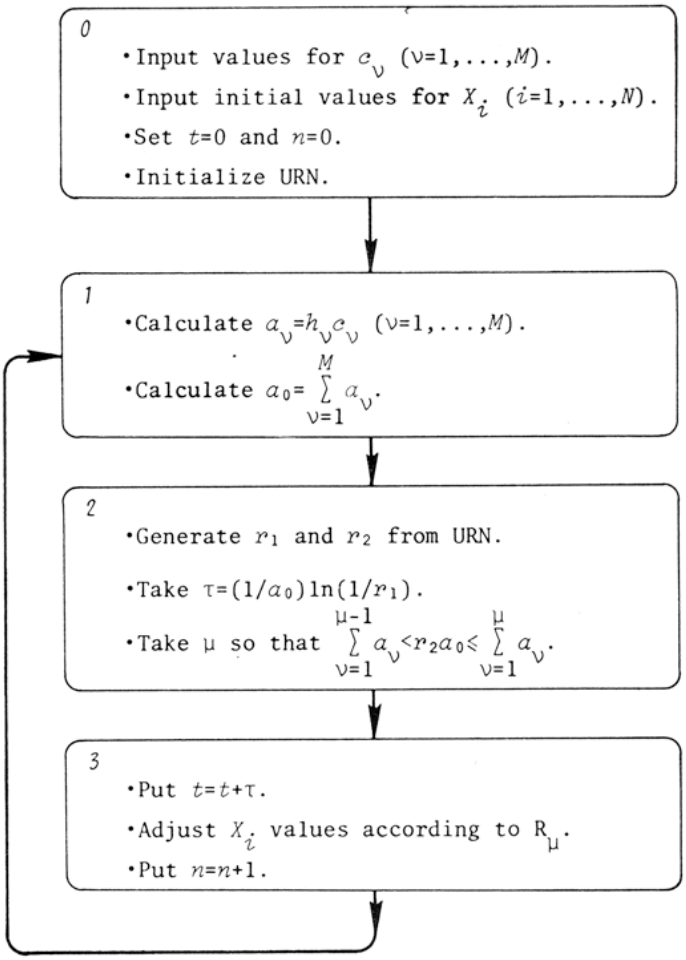

Figure 2.7: Schematic representation of the Gillespie Direct Method.

Chapter 3

Materials & Methods

3.1 The Algorithm

The central motivation behind this paper rests in coupling initiation, elongation and termination which are biophysicaly different events (the first being a three dimensional collision of molecules, the second being the 1D movement along a molecule) into a single consistent framework so as to be able to study how these events take place in stochastic time.

If the algorithm was implemented in a standard Gillespie fashion, every combination of mRNA+Ribosomes should be considered as a species of the system, rendering the system quickly intractable for machines to simulate systems with several mRNAs.

Other methods tend to suffer from having too many variables when the system becomes bigger than a single mRNA strand. Consider in Roussel [18] where the size of the algorithm is dependent on the size of the mRNA. From these points it becomes apparent that developing such an algorithm is not trivial.

Initiation reactions are simple to model in a basic Gillespie way and do not need a particular formulation. However, how to resolve this large number of species that occur during elongation? Consider a ribosome traveling along a strand of mRNA. it is clear that it can only be sitting on one of the 64 codons that make up the natural genetic code. In this sense the ribosome only "sees" one of 64 codons at each elongation step with a particular elongation rate constant K_E associated to it. Therefore, is it possible to adopt the point of view of the bound ribosome to the mRNA to create the set of

reactions that direct the system?

By considering the movement of the ribosome as it travels from a codon x to a codon y, the system becomes limited to 64 reactions for the elongation and termination steps, which are completely independent of the length of the mRNA considered or its sequence.

The main interest behind this approach is the coupling of the two apparently distinct processes of initiation and elongation/termination that actually share the same underlying driving force: mRNA colliding in 3D with ribosomes for initiation, tRNA and factors colliding in 3D with the bound ribosomes for elongation and termination. By being able to merge the two into a single framework, the initiation reactions are no longer decoupled from the elongation steps, resulting in a more unified system.

3.1.1 The Model

Here is a summary of the way the algorithm presented in this work relates to Gillespie's Direct Method Formulation as well as the differences that need to be considered.
What the Gillespie algorithm "sees" is summarized by the reactions below.

$$\begin{aligned} \mathrm{mRNA}_f^m + R_{30s} &\xrightarrow{\mathrm{Ks}_f^m} R_{30s}^m \\ R_{30s}^m &\xrightarrow{\mathrm{Ks}_b^m} \mathrm{mRNA}_f^m + R_{30s} \\ R_{30s}^m + R_{50s} &\xrightarrow{\mathrm{K_2}} R_{\mathrm{AUG}} \\ R_x &\xrightarrow{\mathrm{K_E}(x)} R_y, x = (1, ...64) \end{aligned}$$

In this way, using Gillespie's algorithm, there is only need for calculating $64 + 2 \times m$ propensities, with m being the number of mRNA species to consider.

Using Gillespie's direct approach method, after the calculation of the propensities, the time to the next reaction and what reaction will take place are selected. When the algorithm reaches this step it is broken into three parts: initiation reactions, movement reactions, and termination reactions.

Because we are looking at the number of ribosomes at every codon, one needs to store the information regarding each ribosome's position on the mRNA, the mRNA molecule it is on, and the codon it is currently on. This

allows for the selection of a ribosome to react when an elongation/termination reaction takes place. Similarly, the indices of the free mRNA molecules must be stored, as well as the indices of the Ribosomes and mRNAs that are in the R^m30s form, to know which ones can be bound to the 50s subunit and begin elongation.

Initiation Reactions: These are three reactions that direct initiation.

- Binding of mRNA and 30s subunit (second order reaction):

$$\text{Propensity}: \quad a_1^m = \quad \frac{Ks_f^m}{\text{cellVol.N}_\text{A}} \times [\text{mRNA}_f^m] \times [R_{30s}]$$

  Select (at random) a free mRNA in the system and a free ribosome, Increase R^m30s count by 1 and lower mRNA_f^m and R^{30s} counts by 1. Store the indexes of the mRNA and ribosome that were selected.

- Undinding of mRNA and 30s Subunit (first order reaction):

$$\text{Propensity}: \quad a_2^m = \quad Ks_b^m \times [R_{30s}^m]$$

  Select (at random) an R^m30s and find the index of the mRNA and ribosome that correspond to it. Increase R^{30s} and mRNA_f^m counts by 1 and lower R_{30s}^m count by 1. Remove the ribosome and mRNA indexes so that they are available again.

- Binding of 50s to R_{30s}^m (second order reaction):

$$\text{Propensity}: \quad a_3^m = \quad \frac{K_2}{\text{cellVol.N}_\text{A}} \times [R_{50s}] \times [R_{30s}^m]$$

  Select (at random) an R^m30s and find the index of the mRNA and Ribosome that correspond to it. Find which codon x the ribosome is on. Update the ribosome position to 1, increase R_x count by 1 and store the index of the ribosome. Lower R_{50s} and R_{30s}^m counts by 1. Also, store the position of the Ribosome in an array of size[m][TotalmRNA(m)][mRNALength(m)], called the Occupancy Array to use for checking on ribosome bumping.

Elongation + Termination Reactions: $(\mathrm{R}_x \rightarrow \mathrm{R}_y)$

The movement from one ribosome at a given codon R_x to the next codon R_y in such a way that the next codon y is calculated by $\mathrm{mRNA}_{\mathrm{seq}}[(pos(R_x) + 1] = y$. This limits the system to a set of 64 reactions (one for each codon) for which we can have codon-specific elongation rates (K_E(c), c being one of the 64 possible codons). To these 64 reactions are added a simplified version of the reactions involved in the initiation of transcription, that is, the binding of the 30s Ribosomal Subunit with rate Ks_f, it's eventual unbinding (Rate Ks_b) and the binding of the 50s Ribosomal Subunit (K_2).

Movement along a strand of mRNA is modeled after the Mitarai paper, in which each ribosome is taken to have an occlusion distance of L codons when on the strand. This in turn limits the movement of the ribosome if there is another ribosome in front of it L codons away.

- Movement of a ribosome (first order reaction)

$$\text{Propensity}: \quad a_x = \quad K_E(x) \times [R_x]$$

  Select (at random) an index from the R_x (That is, all the ribosomes that are on codon x). From that index, find the mRNA species, the position of the ribosome on the mRNA and the next codon y.
  Check on the Occupancy array that there is no other ribosome on that particular mRNA closer than L+1 codons from the current ribosome. If there is, then no movement can happen.
  If there is no ribosome occluding the L+1 codon, then movement can take place, which involves: Lowering R_x by 1, increasing R_y by 1, updating the Occupancy matrix and the ribosome indices.

  Also during this step, the algorithm checks for when the codon's position is equal to L, that way it can free the mRNA index again, so that it can receive a new ribosome.

- Termination reaction (first order reaction) If the reaction lands on one of the 3 STOP codons, then termination takes place.
  The propensity is the same as an elongation reaction as they are considered to be ribosome movements.
  Select (at random) an index from the R_x, x here being one of the 3 stop codons. From that index, get the mRNA species, and update the Occupancy matrix. Remove the index of the ribosome, add +1 to

R_{30s} and R_{50s} (unbinding of the ribosomal subunits), and increase the protein count by 1.

The algorithm outputs the following files:

- "Molecules" is a file that contains the number of free mRNAs, free R_{30s}, free R_{50s}, R^m_{50s}, proteins in the system over time, as well as the number of initiations, elongations and terminations that took place during the run.
- "Rhos" contains at each line the total number of ribosomes at a given position on all the mRNA strands for a given mRNA species, over time.
- "Proteins" contains the number of proteins generated by each mRNA individually, over time.
- "Polys" contains the average polysome size of a species m of mRNA, over time.

3.2 Features of the method

As stated above, this formulation captures the coupling between the elongation events (through tRNA competition) to the initiation events. The ribosome movement is independent on an mRNA strand but the choice of having an elongation reaction is dependent on the propensities of each other reaction in the system as well, thanks to Gillespie.

Scaling: One of the most interesting aspects of this algorithm is that the number of reactions does not scale with the number of molecules of in the system. This way, the data recovery remains very much identical no matter what system and how many molecules are being studied. The system does increase by 3 reactions each time a new mRNA species is added, to account for varying Ki^m_f, Ki^m_b and K_2.

Easily Expandable: The Gillespie backbone allows for very easy addition of any number of reactions, such as different termination rates per mRNA, or modelling initiation in a more complete way, taking into account other intermediates, as seen in figure 2.1.

Fast: For an mRNA strand of 144 codons, 100'000 ribosomes and 100 mRNAs, the algorithm takes less than a minute to generate one trajectory where 10'000 proteins are produced,

3.2.1 Sample Trajectory

In the annex figure 9.2 is an example trajectory of a system of 1000 identical LacZ Operon mRNAs competing for 18'000 Ribosomes, with parameters borrowed from Mitarai *et al.* The algorithm was run until 10'000 proteins were produced, and the results were viewed using the R tool for Statistical Computing [1], with the elongation rates from the Mitarai paper (See figure 9.3) in the Annex. $\mathrm{Ks}_f = 3.10^8 s^{-1} Mol^{-1}$, $\mathrm{Ks}_b = 10 s^{-1}$, $\mathrm{K}_2 = 3.10^8 s^{-1}.Mol^{-1}$.

[1] www.r-project.org

Chapter 4

Results

Most of the results in this report deal with the recovery of the H-R model from [6], because the size of the system (100 mRNAs of length 144 codons and 100'000) Ribosomes allowed for fast simulations. A wide range of parameters could be studied in this way. The mRNA that was input into the system is the human α-globin gene that H-R used in their model due to the experimental data available from its recombinant expression in *E.coli* .

Also, for these results, the model was further simplifed by considering the binding of the ribosome to the mRNA as being single-step. This implies that $\mathrm{Ks}_b = 0s^{-1}$ and $\mathrm{K}_2 \gg \mathrm{Ks}_f$. Running a series of trajectories where the Ks_f was varied from $0s^{-1}$ to $1.10^4 s^{-1}.Mol^{-1}$ by increments of $200s^{-1}.Mol^{-1}$ and the K_T (The termination rate) from $0s^{-1}$ to $0.5s^{-1}$ by increments of $0.002s^{-1}$ yields the surface plots seen in figure 4.1. Each trajectory was run at steady state until 10'000 proteins were produced (On average, 100 per mRNA), sampling the data every 500^{th} iteration of the algorithm. The elongation rates K_E were fixed to $1.0\mathrm{s}^{-1}$. Figure 4.1 represents about 20 Gigabytes of data, and each point in the surface plot is read from the files output by the algorithm (4 files generated for each trajectory).

4.1 Recovering the H-R Model

Using figure 4.1 we can then recover plots for varying Ks_f and a fixed K_T, which gives us plots similar to the Heinrich-Rappoport model (Figure 4.2). The results are for only a fixed K_E value of $1.0\mathrm{s}^{-1} = 60\mathrm{min}^{-1}$. Unlike the H-R model, there is no abrupt jump in polysome sizes. The behavior of the protein production curves matches the observed behavior of the H-R model, for similar values of K_T (black and red curves vs curves a and b from H-R).

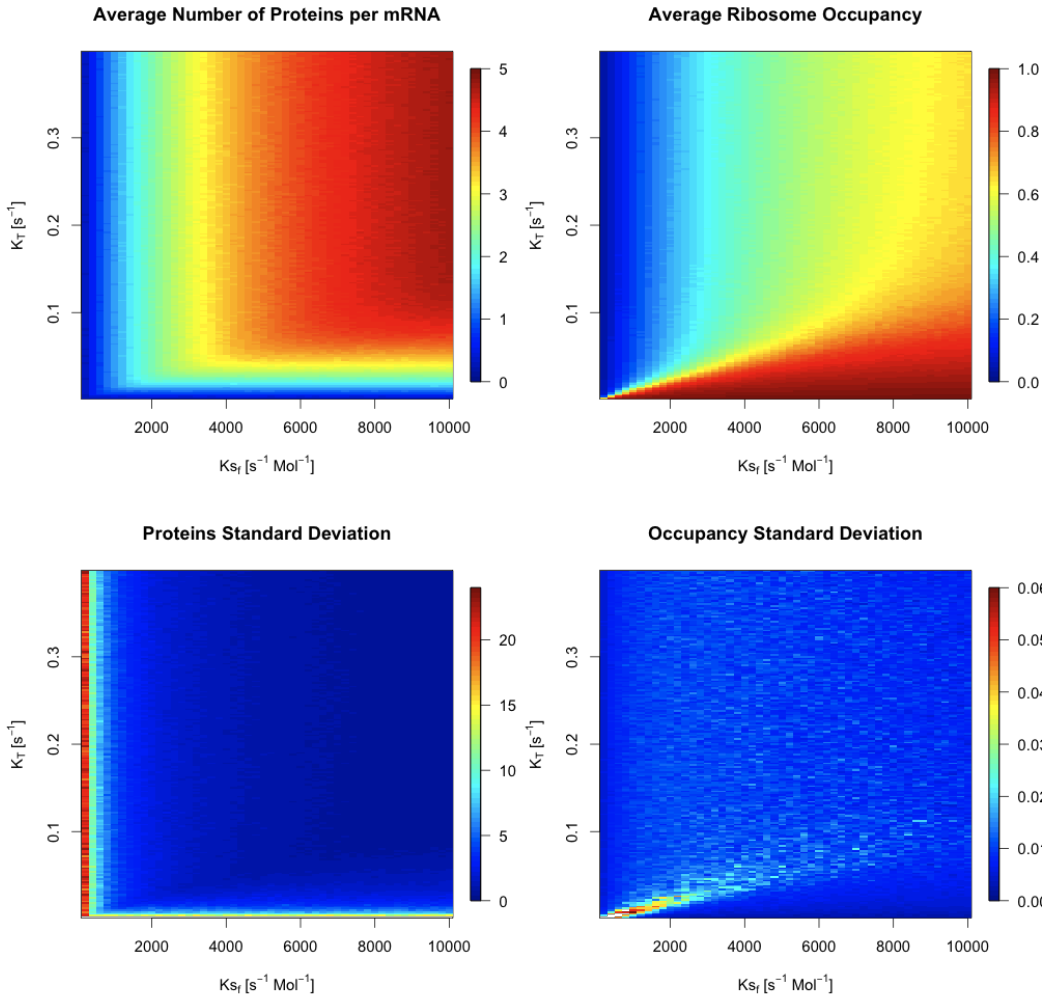

Figure 4.1: Surface plot obtained from runs of the stochastic translation algorithm. Top Right: Average number of proteins per mRNA in Proteins.s^{-1}, Top Left: Average polysome size per mRNA. Bottom Left: Standard deviation in protein production for each (Ks_f, K_T) pair. Bottom Left: Idem for polysome sizes.

We also observe a direct anti-correlation between polysome size and protein production rate.

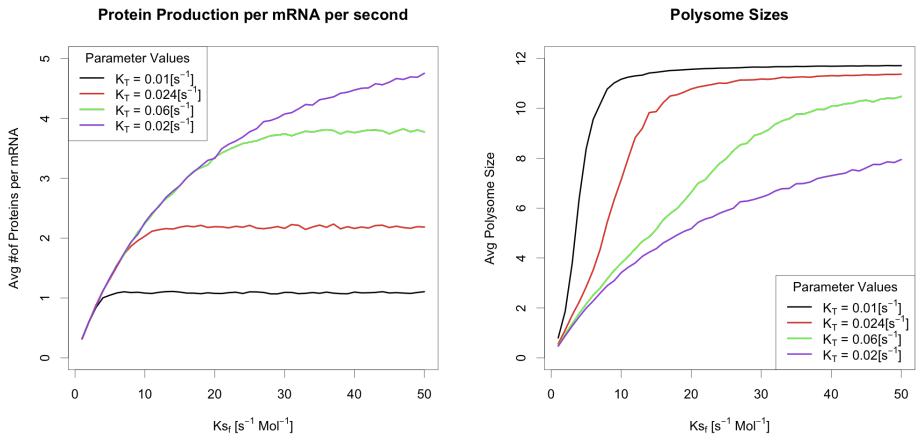

Figure 4.2: By following along a given k_T on the surface plots, we can trace, for a fixed K_E, the variation in polysome size as well as the protein production of the considered mRNA. The curve at $K_T = 0.2s^{-1}$ is the maximum protein production rate obtainable for any value of $K_T > 0.2s^{-1}$.

4.2 Recovering the Z-H Model

Again using the surface plot data, by plotting the ribosome density on the x axis and the protein production rate on the y axis, first along a fixed K_T, then lowering the K_T value, plots close to the Z-H model can be recovered. Figure 4.3 was obtained by initially setting $K_T = 0.3s^{-1}$ and varying Ks_f from 200 to 10'000$s^{-1}Mol^{-1}$. At $Ks_f = 10'000s^{-1}Mol^{-1}$, K_T was lowered from 0.3 to $0.002s^{-1}$.

The result compares directly to the plot from the Z-H model that was presented in figure 2.4. The accumulation of points at the top of the plot show many similar (Ks_f, K_T) to be responsible for the same production rates. The maximum production rate is near to 5 proteins per second at a density of about 0.65. Considering that the influence of the elongation rate was not studied in this context, one cannot conclude directly that this is indeed the highest production rate available to this system.

4.3 Recovering the Queueing

A simulation was run, mimicking the parameters used by Mitarai et al, where the queueing that was observed in their paper is recovered using the new algorithm (Figure 4.4). Here the mRNA used was the lacZ operon to which

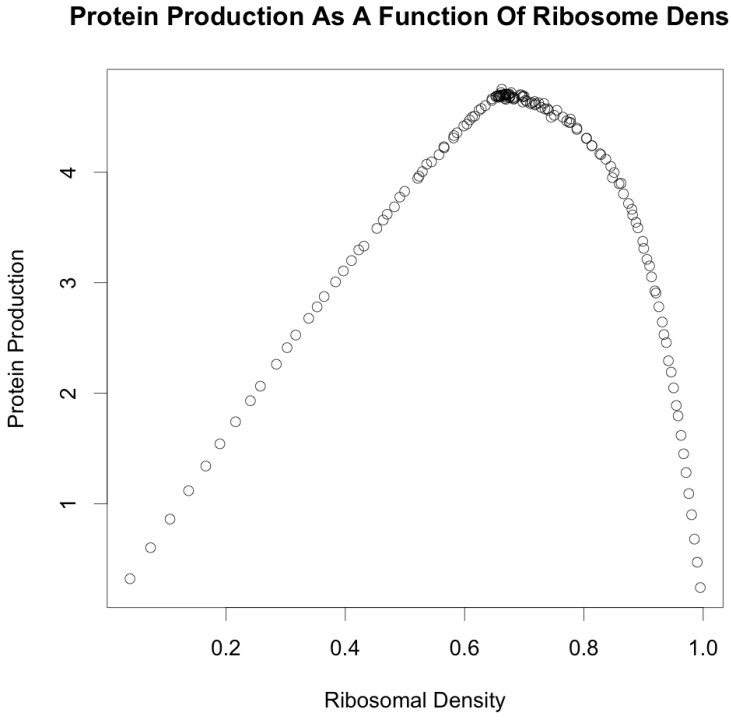

Figure 4.3: Specific protein production as a function of polysome size. The large number of dots around the top represents points where varying the Ks_f or K_T has little impact on the polysome size for a given protein production rate or vice-versa. K_E=1.0s^{-1}.

the sequence CGA$(GAG)_8$CCG was inserted 6 times at nucleotide position 2779. GAG is considered a slow codon by the Mitarai paper (See Annex), so adding a long series of these should have an effect on the ribosome distributions on the mRNA.
The simulation was run for 1 mRNA at a very high ribosome numbers (18'000) so as to ensure that the binding of ribosome to mRNA was not the limiting step driving the system. A variable was created to record the positions over time of each ribosome that had irreversibly bound to the mRNA.

As in in the Mitarai paper, the queueing is very apparent at the end of the sequence, where the insert was added, as shown by the ribosome translocation rates becoming smaller, because ribosomes are spending more time at the slow codons (more important negative slopes on the graph).

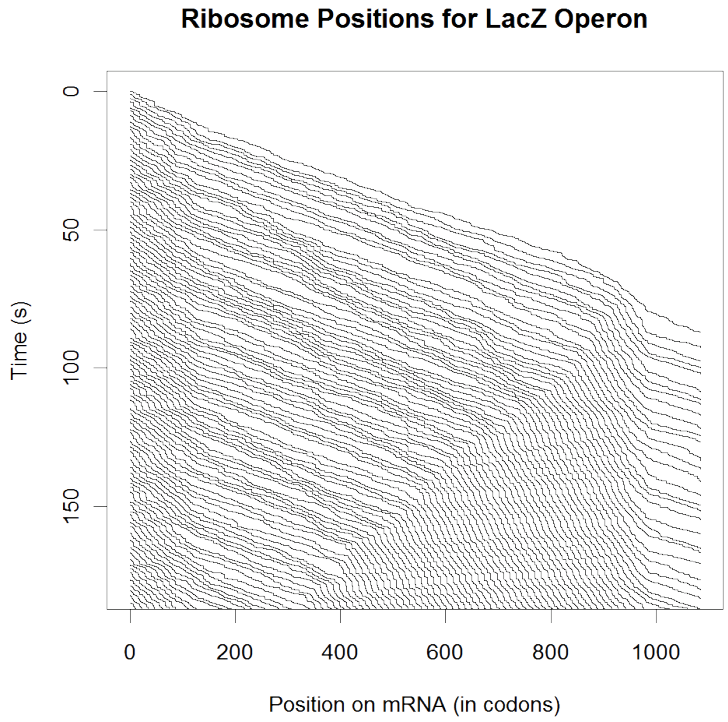

Figure 4.4: Queueing observed in a high Ribosome concentration with the following parameters: $Ks_f = 3.10^8 s^{-1}.Mol^{-1}$, $Ks_b = 5s^{-1}$, $K_2 = 10^8 s^{-1}.Mol^{-1}$, for 1 mRNA in the presence of 18'000 Ribosomes. The algorithm was run until 50 proteins were produced. The mRNA used is a lacZ Operon with an added duplicate of slow codons after nucleotide 2779.

4.4 Implicit Noise, Expanded Results

4.4.1 Looking at the noise in the H-R Model

Noise polysome noise levels were plotted for $K_T = 0.02s^{-1}$ as a function of Ks_f and as a function of the ribosome density. These values show two zones where the polysome noise is important. One for low values of Ks_f or initiation limited translation, and a second at a ribosome density of around 0.6. Considering that the maximum polysome size in the system is 12 (mRNA length divided by ribosome occlusion distance), this corresponds to a polysome size of around 7.2 which is the zone where we would expect the H-R model to leap in polysome size, for the same value of K_T.

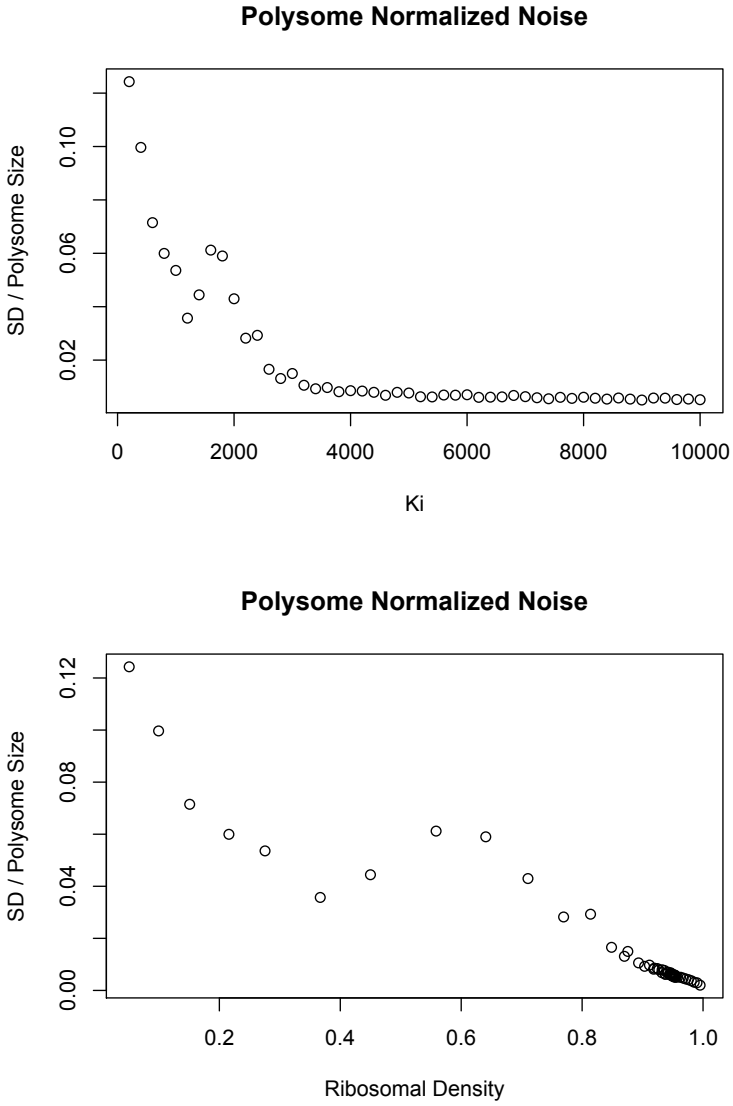

Figure 4.5: Top: Normalized standard deviations of polysome noise for increasing Ks_f values. The system is very noisy for low values of Ks_f ($K_T = 0.02.s^{-1}$ Bottom: Plotting normalized polysome size standard deviation versus ribosomal density shows a bump in polysome noise centered at around 0.6 density.

4.4.2 Expanding on the Z-H Model

When looking at the standard deviation surfaces, ploting the normalized standard deviation of the protein production rate with respect to polysome

size results in figure 4.6.

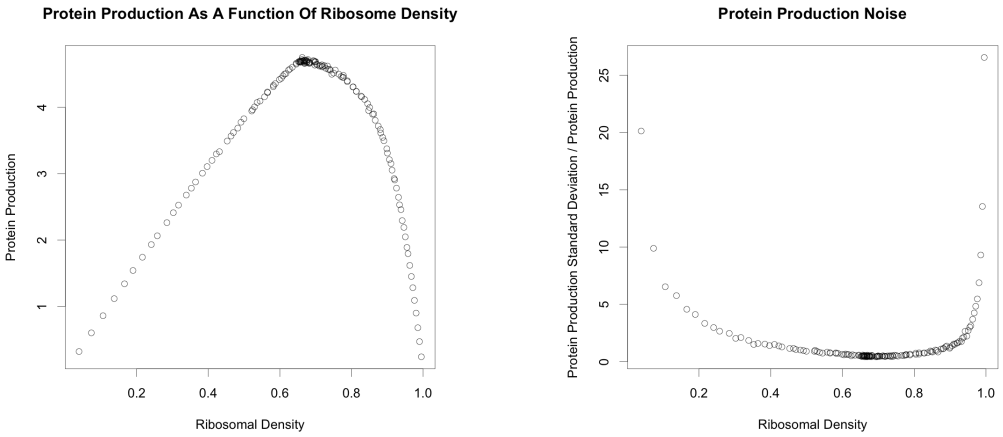

Figure 4.6: Normalized standard deviation of protein production plotted against polysome size. As the system reaches an optimal polysome size for maximum protein translation, the uncertainty on the number of proteins being produced goes down.

This result suggests that optimal protein production rates due to polysome self-organization also work towards minimizing the uncertainty in the output of the system.

It is also possible to recover the ribosome distributions for each point of figure 4.3 and verify the H-Z conclusion that the maximum protein production rate occurs when the ribosome distributions along the mRNA are still uniform. Figure 4.7 illustrates four distributions selected at 5 different positions of figure 4.3, going from low polysome size, reaching optimal protein translation, at maximum protein translation, at high ribosome density and at maximum ribosome density.

As the system nears the maximum ribosome density, the ribosome distributions become less uniform, because of ribosomes are forced to stay longer at certain codons. The phenomenon starts at the end of the mRNA, due to the kinetics being termination limited.

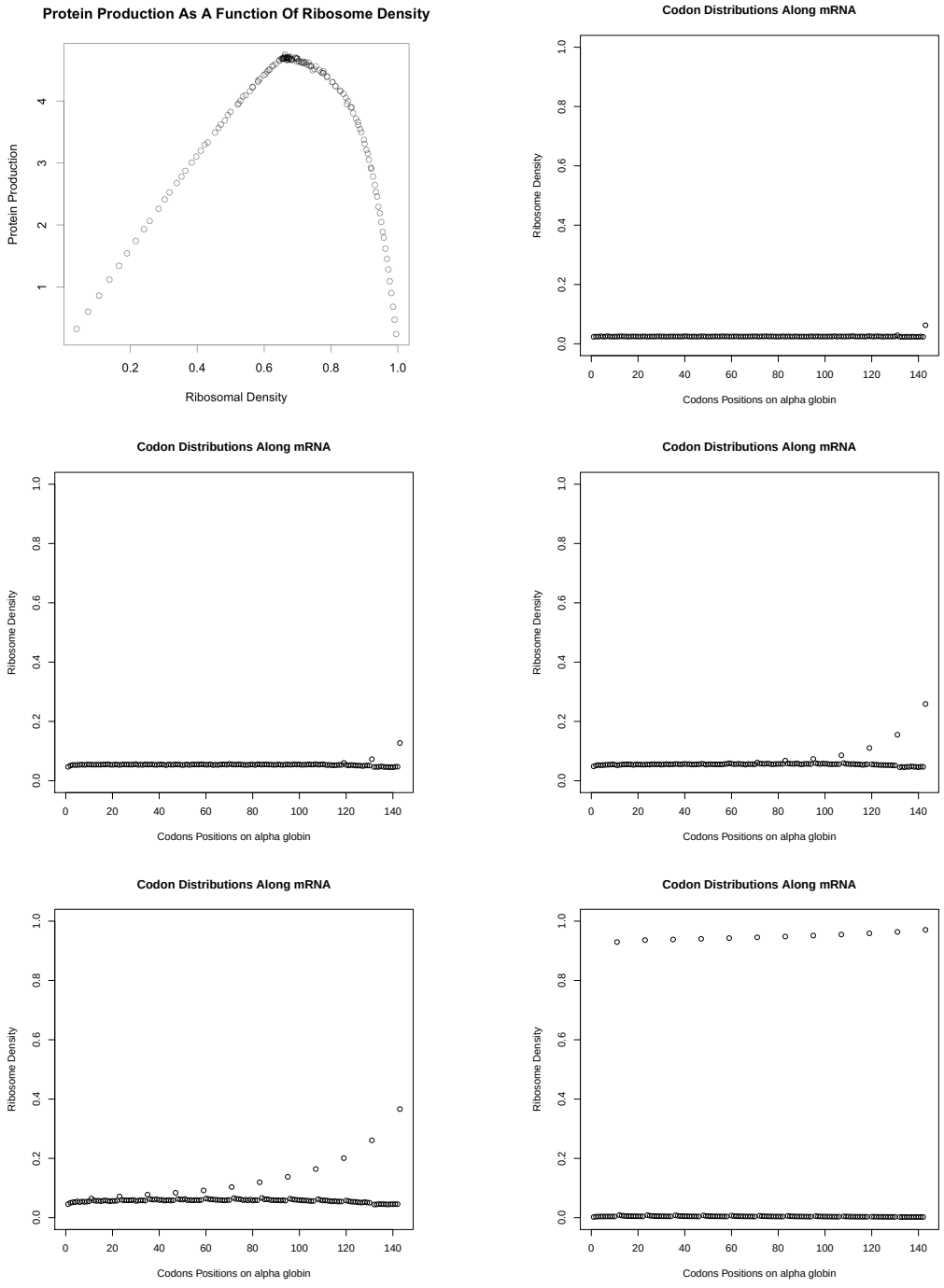

Figure 4.7: Looking at the ribosome densities for different values of Ks_f and K_T. Going from left to right: Copy of figure 4.3, $\mathrm{Ks}_f = 2000 s^{-1} Mol^{-1}$, $\mathrm{K}_T = 0.3 s^{-1}$, Middle: $\mathrm{Ks}_f = 10'000 \mathrm{s}^{-1} Mol^{-1}$, $\mathrm{K}_T = 0.3 s^{-1}$, and, $\mathrm{Ks}_f = 10'000^{-1} Mol^{-1}$, $\mathrm{K}_T = 0.15 s^{-1}$, Bottom: $\mathrm{Ks}_f = 10'000 s^{-1} Mol^{-1}$, $\mathrm{K}_T = 0.1 s^{-1}$ and $\mathrm{Ks}_f = 10'000^{-1} Mol^{-1}$ and $\mathrm{K}_T = 0.002 s^{-1}$

Chapter 5

Discussion

Throughout the examples, the validity of the stochastic translation algorithm is demonstrated, the obtained data matches well with the existing data from validated models. By looking at noise distributions of protein production, the algorithm offers both an elegant result as well as a great window of opportunity for further work in this direction.

Discussion for the H-R results: The recovery of the Heinrich-Rapoport model seems, at first not complete . The "jumps" in polysome sizes were not observed in figure 2.2. The parameters chosen did not exactly match the H-R polysome size curves either. However, it is possible that this jump behavior arises from bistability of the parameters at high ribosome densities. In that case looking at the distributions of polysomes in the system offers a better understanding of the results. If a bistable behavior exists, then there should be two distinct distributions of polysome sizes, one working at a high regime and the other at a lower one, visible in the noise distributions as an increase of the standard deviation in polysome sizes. Because we only look at the average polysome sizes, it is entirely possible that this behavior is smoothed out in the recovered H-R curves. A proof that this behavior is taking place can be seen in figure 4.5. The bump in polysome standard deviation noise shows that at a regime between $\mathrm{Ks}_f = [1700 - 2100]\mathrm{s}^{-1}\mathrm{Mol}^{-1}$ and a ribosome density of about 0.6 (Polysome size=7.2) for $\mathrm{K}_T = 0.02\mathrm{s}^{-1}$ (corresponding to the red curve in the H-R plots and close to curve (b) in the H-R data) appears where the leap in polysome size is expected. This suggests that there is a change of phase, where the noise of the system becomes important before switching to a what could be called a crystallized state. This behavior, which was not captured when averaging the polysome sizes over time for the H-R model, is illustrated by figure HR-polysomeNoise

So the algorithm did correctly recover the H-R model, and gave results that could be matched properly to the expected behaviors. Looking at a polysome distribution could help determine more precisely the nature of this behavior, by showing the presence of multiple polysome size sub-populations.

Discussion for the Z-H results: The ribosome density vs. protein production plots were recovered in good agreement with the Zouridis-Hatzimanikatis data, even when considering constant elongation rates. Overlapping the protein noise distributions with respect to ribosome density showed that as the system reaches a maximal protein production rate, the noise levels are reduced, suggesting that at steady state this is the optimal behavior for protein production. Typical textbook data suggests that 80% of ribosomes are active at any time, so about 14'400 ribosomes are on the 4'000 mRNAs that are on average 1'100 bases long, or 360 codons. If we consider the 14'400 ribosomes, there are on average 4 ribosomes available per mRNA. With an occlusion distance of 12 codons, there can be a maximum of 30 ribosomes on an mRNA molecule. So the average occupancy is around 0.15. Figure 4.5 shows that low noise levels start at around a density of 0.3, which, considering the very broad variations in obtaining the textbook data, confirms, in terms of noise levels that the concentrations automatically allow for this low noise level. An interesting follow-up result would involve the reproduction of figure 4.5 with the average parameter values in *E.coli* to further validate this result.
One conclusion of the Z-H model states that the maximum translation rate is found for Ks_f,K_T pairs for which the ribosome distributions are uniform along the length of the mRNA. Figure 4.6 shows indeed that, as we get closer to a ribosome density of 1.0, the ribosome distributions become less uniform, reaching a point where the mRNA is blocked by 12 queueing ribosomes (Figure 4.6 , bottom right). Interestingly, the noise is not minimal for ribosome densities lower than 0.6, showing that perhaps a certain density of the system is needed to reduce the uncertainty. Just enough to avoid queueing, and interesting results should appear when variable translation rates are included into the model.

Queueing results: For the slow lacZ operon, the algorithm recovers the queueing behavior properly when considering the variable elongation rates suggested by Mitarai *et al.* The system models the behavior of ribosome queueing explicitely and gives very clean results when viewing individual mRNAs, but the study on different mRNA populations, and the polysome

distributions varying over time could give rise to very interesting behaviors to study, that were, unfortunately not carried out during this project.

Chapter 6

Future Outlook

The work presented in this Masters is intended to represent a validation of the use of this stochastic translation algorithm, which offers the possibility to answer an important number of interesting questions, as well as modified to encompass the ever-growing amount of knowledge that exists on this subject, and perhaps bring interesting answers and new emergent behaviors.

Many Species: The results and figures shown in this report deal only with a single mRNA species (either the α-globin gene or the LacZ operon), but the algorithm is capable of generating a pool of multiple mRNA species, allowing for the study of mRNA competition for resources, like ribosome availabilities, that could offer some interesting insights as supplemental forms of protein translation regulation.

Using The Variable Codon Rates: For the most part, a fixed elongation rate was used throughout this work, and even when multiple elongation rates were used, as in the queueing example, it only involved either slow, normal or fast codons, without further development. In a second paper by Zouridis & Hatzimanikatis [21], the codon translation rates are calculated for each codon, adding a supplemental layer to their model, and bringing it closer to the physical reality of the system. In the case of the stochastic translation algorithm, one could envision deriving rate expressions for the different codons as well, so as to further study their effect on protein translation and ribosome density distributions.

More Complex Initiation: As shown in the introduction, prokaryotic initiation of translation has several more players than just the mRNA and the Ribosomal subunits, in cases of initiation-driven protein translation. Adding

these supplementary species can help discover where the rate-limiting reactions are, and how this gives rise to the kinetic rate constants that were used in this project and more importantly the distributions of the molecules.

Stalling: In a recent article by Jin-Der Wen et al [19], an experiment is described where a single codon is followed as it translates a strand of mRNA using optical tweezers to "see" the opening of the mRNA hairpin as the ribosome opens it. One of the conclusions is that ribosome stalling occurs, which could be modeled into the system. For this one would have to find a criterion for stalling and check whether this addition modifies the output of the algorithm.

Consider Steric Hindrance in the elongation rates: As a ribosome gets closer to another, the electrostatic forces occurring in the proximity should modify the affinity of tRNA for the A site, reducing it as the ribosomes get closer. By coding this aspect explicitly into the algorithm, a more realistic system can be obtained. In practice, this would involve generating a table of correction values for the K_E for each ribosome that is bound, and modify it each time a ribosome moves. Then, instead of choosing a ribosome index at random, (uniformly), the selection can be done by using the correction values as a weight matrix for selecting the ribosomes.

mRNA Birth, Decay and energy expenditure: Several papers and, in this case, the Mitarai paper explain how maximum protein synthesis with respect to resource allocation is maximized for small-lifespan mRNAs. This puts in question the concept of studying these systems with fixed non-varying mRNA numbers and at steady state. Adding mRNA birth and decay explicitly might bring about new behaviors and modes that could not be considered within this work. Because the algorithm already records the number of initiation, elongation and termination events, these can be coupled to an energy usage function that would calculate the ATPs that are consumed at each time step, providing the same type of data that the Mitarai paper was able to produce.

Variable kI based on ΔG or other strategies: There is a large body of data dealing with the modeling of transcription initiation in prokaryotes: Models based on the Shine-Dalgamo sequence or even others showing how changes in the codon +1 to +5 upstream of the S-D sequence could yield 10-fold changes in protein expression. One approach showed that the Gibbs free energy was well correlated to initiation rates. It would be interesting to

code this into the algorithm, providing a dynamic Ks_f value, dependent on the conformation of the 5' end of the mRNA.

Explicitly Code tRNA Competition: The second paper by Zouridis deals with the competition of tRNAs for binding to the Ribosomal A site, showing it to be the rate-limiting step in the elongation cycle [21]. It should be possible to code this competition explicitly by considering a system of 64 $\times$ 64 reactions, where each tRNA has a certain probability of being able to reach the Ribosomal A site, and then unbind.

Noise Propagation Through The System: An aspect which is interesting to consider is that the 1D movement of the Ribosome along the mRNA could act as a low-pass filter, removing the high-frequency noise that arises from rapid binding an unbinding of the 30s subunit, or any other element in the initiation cascade. By using a more complete initiation model, one could then consider outputting "virtual" proteins at given lengths of the mRNA and see how their distributions change as the ribosomes move towards the 3' end. Variable elongation rates also have a role to play, as was shown in the Mitarai paper and the Zouridis paper.

Other Gillespie implementations: The direct method of the Gillespie algorithm was sufficient for the purposes of this work, but improvement of the speed of the algorithm can be obtained by using Gillespie's Next Reaction method, which reduces the need for pseudorandom number generation, removing a certain calculation load from the processor. Code optimization to avoid using complex or poorly efficient functions for indexing can also further improve the speed.

Sensitivity Analysis: a very important aspect of systems biology deals with sensitivity analysis, which are mathematical tools that allow for the identification of the sources of variation or uncertainty in a mathematical model. There is a large body of sensitivity analysis tools available for deterministic continuous systems, as well as literature for stochastic models [5]. Adding sensitivity analysis to the model suggested would help gain a more mathematically sound and robust understanding regarding the interplay and influence of each reactant and product in the system.

Friendly Interface: As of now, the algorithm is a simple program compiled each time new parameters are input. A clean code that would allow for interested scientific collaborators to generate their own simulations in an

easy-enough-to-use fashion would help both boost the interest in this algorithm as well as allow for a faster and quickerr selection of parameters for simulation.

Chapter 7

Conclusion

For this project, the goal was to produce a stochastic translation algorithm that consisted of a unified framework in which the elongation and termination reactions are described in terms of the ribosomal point of view, that is, looking at transitions of the ribosomes from one codon to the next rather than the evolution of all the mRNA / ribosome species combinations that exist in the system. These reactions can then be directly coupled to the initiation reactions that can be modeled using Gillespie's algorithm.

Testing the stochastic translation algorithm against the discrete queueing paper by Mitarai et al, which was the starting point of this project, showed that through the Gillespie formulation, the queueing behavior observed in their results could be replicated, with the added possibility to look at noise distributions of the ribosomes and the polysomes as they move towards the 3' end of the mRNA (Data not shown here).

Borrowing from the Heinrich and Rapoport model, the algorithm was capable of replicating the results of polysome size pileup. The jump behavior is explained through the noise distributions of polysome sizes along the ribosomal density to show that it is reminding of a phase change in the system, where the polysome sizes switch from low density to higher density, like crystallization in physical systems. A follow-up work should look at the polysome size distributions for the mRNA population, to see if at the regime switching point, we can observe different sub-populations in polysome size.

The conclusions of the Zouridis-Hatzimanikatis model presented in the introduction were recovered, namely that increasing polysome sizes leads to increased protein translation rate, up to a maximum when the kinetics become termination limited. Also, maximum production rates occur while the

ribosome density remains uniform. Also, the capacity at viewing noise in the stochastic model was illustrated by showing how protein noise distribution as a function of ribosome density reached a minimum when at maximum protein production. The plateau of noise is also reached at around a density of 0.3, corresponding to the average ribosomal occupancy in *E.coli* .

Ribosomal queueing was visible when using the parameters from Mitarai et al's Kinetic Monte Carlo simulation, showing the equivalence in the formulation of the algorithm.

The validation of the algorithm can be considered successful and opens the possibility to answer many questions regarding the noise distributions in the system when considering multiple mRNA species and different elongation rate constants for each codon, based, for example on tRNA availability and competition. Sensitivity analysis of these systems should provide valuable information about the influence of the different parameters and species to the global behavior of the system. The simple implementation of the algorithm makes it easily expandable to account for many other parameters and provides a good starting point for research in stochastic translation.

Chapter 8

Acknowledgments

Through the course of this four months masters project, I have managed to learn about the amazing field of systems biology, under the tutoring of Prof. Hatzimanikatis who was always available to listen and give great advice, and Jan Overney followed me through the project, directing me to the right publications and always available to discuss issues. They have helped me reach a sufficient understanding in the field and the tools available to make this project a success.
Many thanks to Elizabeth Brunk for listening and finding an interest in the project and offering valuable input, to Emily Hammes for helping me parallelize the simulations, gaining an 8-fold gain in time, to Luis Mier y Teran for publications and discussions regarding previous work, to Pooya Pakzad, for always being there when I needed to take a break, to Fabien Roland for being supportive and kind during these four months of work.
Thank you to Jasper Bedaux[1] for his implementation of the Mersenne Twister pseudo-random number generator C++ code that was used in the algorithm
Finally thank you to my parents, Ruth & Jacques Burri, for financing my studies and making it possible for me study what I consider to be one of the most interesting fields in research. I have found my way.

[1] http://www.bedaux.net/mtrand/

Bibliography

[1] D Bratsun, D Volfson, LS Tsimring, and J Hasty. Delay-induced stochastic oscillations in gene regulation. *P Natl Acad Sci Usa*, 102(41):14593–14598, Jan 2005. masters.

[2] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25):2340–2361, Dec 1977.

[3] H Goodarzi, HA Nejad, and N Torabi. On the optimality of the genetic code, with the consideration of termination codons, Jan 2004.

[4] V Grubelnik. Signal amplification in biological and electrical engineering systemsuniversal role of cascades. *Biophysical Chemistry*, 143(3):132–138, Aug 2009.

[5] R Gunawan. Sensitivity analysis of discrete stochastic systems. *Biophys J*, 88(4):2530–2540, Apr 2005.

[6] R HEINRICH and TA RAPOPORT. Mathematical-modeling of translation of messenger-rna in eukaryotes - steady-states, time-dependent processes and application to reticulocytes. *J Theor Biol*, 86(2):279–313, Jan 1980. masters.

[7] S Hooshangi, S Thiberge, and R Weiss. Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. *P Natl Acad Sci Usa*, 102(10):3581–3586, Jan 2005.

[8] YN Kaznessis. Multi-scale models for gene network engineering, Jan 2006.

[9] T Kepler and T Elston. Stochasticity in transcriptional regulation: Origins, consequences, and mathematical representations. *Biophys J*, Jan 2001.

[10] Pat S Lee, Leah B Shaw, Leila H Choe, Amit Mehra, Vassily Hatzimanikatis, and Kelvin H Lee. Insights into the relation between mrna and protein expression patterns: ii. experimental observations in-escherichia coli. *Biotechnol. Bioeng.*, 84(7):834–841, Dec 2003.

[11] H McAdams and A Arkin. It'sa noisy business! genetic regulation at the nanomolar scale. *Trends in Genetics*, Jan 1999.

[12] HH McAdams and A Arkin. Stochastic mechanisms in gene expression. *P Natl Acad Sci Usa*, 94(3):814–819, Jan 1997.

[13] Amit Mehra and Vassily Hatzimanikatis. An algorithmic framework for genome-wide modeling and analysis of translation networks. *Biophys J*, 90(4):1136–1146, Feb 2006.

[14] PB MILANOV, PS KENDEROV, and OC IVANOV. On the optimality of the genetic-code, Jan 1986.

[15] N Mitarai, K Sneppen, and S Pedersen. Ribosome collisions and translation efficiency: Optimization by codon usage and mrna destabilization. *Journal of Molecular Biology*, 382(1):236–245, Sep 2008.

[16] HS Najafabadi, H Goodarzi, and N Torabi. Optimality of codon usage in escherichia coli due to load minimization. *J Theor Biol*, 237(2):203–209, Jan 2005.

[17] CV Rao, DM Wolf, and AP Arkin. Control, exploitation and tolerance of intracellular noise, Jan 2002.

[18] Marc R Roussel and Rui Zhu. Stochastic kinetics description of a simple transcription model. *Bull. Math. Biol.*, 68(7):1681–1713, Sep 2006.

[19] Jin-Der Wen, Laura Lancaster, Courtney Hodges, Ana-Carolina Zeri, Shige H Yoshimura, Harry F Noller, Carlos Bustamante, and Ignacio Tinoco. Following translation by single ribosomes one codon at a time. *Nature*, 452(7187):598–603, Apr 2008.

[20] Hermioni Zouridis and Vassily Hatzimanikatis. A model for protein translation: Polysome self-organization leads to maximum protein synthesis rates. *Biophys J*, 92(3):717–730, Jan 2007. masters.

[21] Hermioni Zouridis and Vassily Hatzimanikatis. Effects of codon distributions and trna competition on protein translation. *Biophys J*, 95(3):1018–1033, Jan 2008.

Chapter 9

Annex

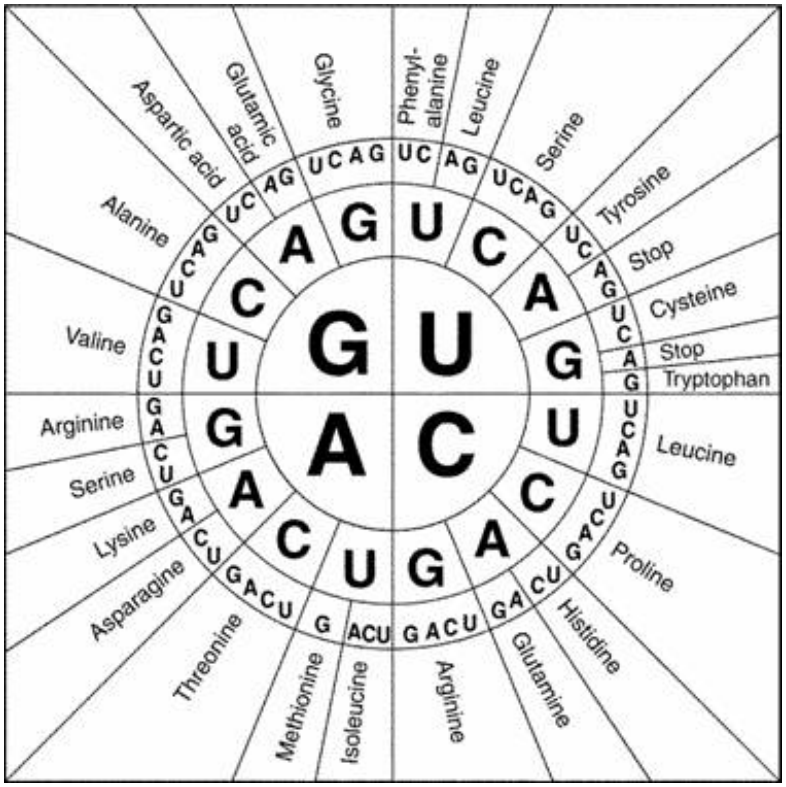

Figure 9.1: Natural Genetic Code: Codons and their corresponding amino-acids.

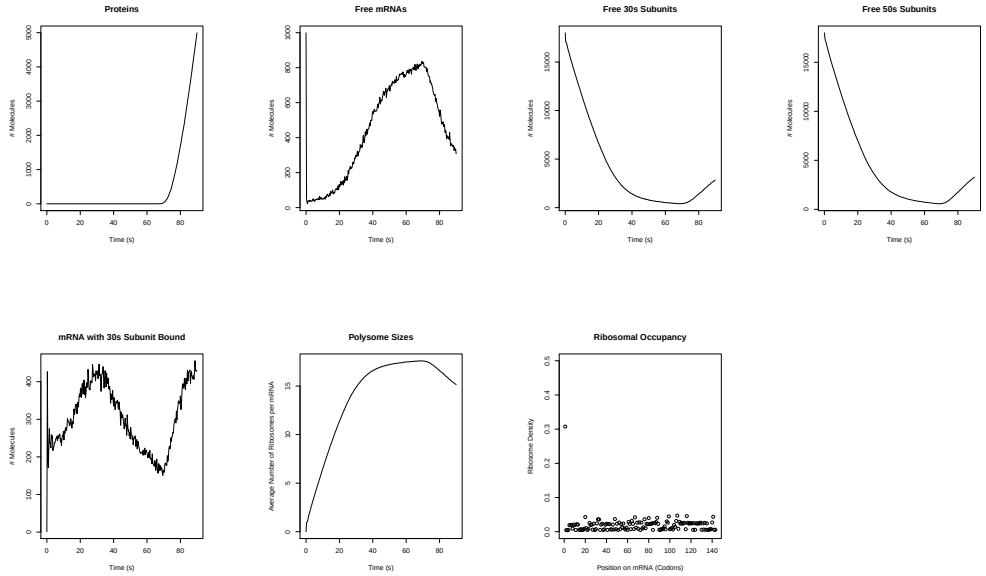

Figure 9.2: Example output trajectories of the stochastic translation algorithm for a system of 1000 lacZ opeon mRNAs run until 5'000 proteins were output. The ribosomal occupancy is an average over time but the system outputs the occupancy at each desired time step to a time graph can be generated.

TTT	B	TCT	A	TAT	B	TGT	B
TTC	A	TCC	A	TAC	A	TGC	A
TTA	B	TCA	B	TAA	–	TGA	–
TTG	B	TCG	B	TAG	–	TGG	A
CTT	B	CCT	C	CAT	B	CGT	A
CTC	B	CCC	C	CAC	A	CGC	A
CTA	B	CCA	C	CAA	B	CGA	C (4.2)
CTG	A	CCG	B (7.0)	CAG	A	CGG	C
ATT	A	ACT	A	AAT	B	AGT	B
ATC	A	ACC	A	AAC	A	AGC	B
ATA	C	ACA	B	AAA	A	AGA	C
ATG	A	ACG	B	AAG	B	AGG	C
GTT	A	GCT	A	GAT	B	GGT	A
GTC	B	GCC	B	GAC	A	GGC	A
GTA	A	GCA	A	GAA	A (35.0)	GGA	C
GTG	A	GCG	B	GAG	B (7.5)	GGG	B

Figure 9.3: Codon transcription rates selected by Mitarai *et al.* for their model. For the purposes of this masters, $A = 35s^{-1}$, $B = 8.0s^{-1}$ and $C = 4.5s^{-1}$