Stochastic Models for mRNA Transcription

Beth Reed

A thesis presented in partial fulfillment of
the requirements for the degree of
Bachelor of Science

Department of Physics and Engineering
Washington and Lee University
Lexington, VA, USA
April 15th, 2022

Abstract

Gene regulation is essential for the diversity and maintenance of life. The first step in protein production, and therefore the first step in gene regulation activities, is transcription. This thesis will outline and analyze three different stochastic models of mRNA transcription. Stochastic models capture the randomness activity on a microscopic level within cells to then predict macroscopic behavior. The stochastic models presented here begin with a one-state model and increase in complexity to three state and multi-state, models. Future goals include increasing model complexity to more accurately capture biological phenomena, and incorporating real biological rates of mRNA birth and death as well as rates of gene activation and inactivation.

Acknowledgments

Firstly, I would like to thank my God, my family, and my friends who have become like family. Without them I would not have had nearly as much joy or support when undergoing this theses. Mom and Dad, you taught me how to love others and where to find peace in this hectic world, which are gifts that deserve an endless amount of thanks.

Secondly, I would like to thank the Physics and Engineering Department of Washington and Lee University and the amazing professors that went into my undergraduate education. In this department I found a collection of passionate, empathetic, and encouraging teachers. While classes were not always easy, they were rewarding, and for that I will be forever appreciative. I would like to give special thanks to Dr. Irina Mazilu, Dr. Dan Mazilu, and Dr. Laurentiu Stoleriu for the unparalleled contributions they made to this thesis.

Lastly, my overwhelming gratitude is extended toward Dr. Irina Mazilu. I can confidently say she is the reason I am a physics major and the reason this thesis exists. The opportunities I have seized at this university, and my future career, would never have come to fruition without the endless support, mentor-ship, and care she devoted to me and selflessly gives to so many others.

Contents

Chapter 1

Introduction

Statistical Physics is a field dedicated to understanding the macroscopic behavior of complex systems comprised of many constituents. Equilibrium statistical physics specialized in studying systems that move towards an equilibrium state. However, life is a non-equilibrium system and the methods of classical statistical physics do not apply, so new methods of study are needed.

Non-equilibrium statistical physics is particularly adept at studying biological processes. Very few systems that occur in nature follow equilibrium statistical physics as the majority of biological processes are irreversible processes.. Biologists and biochemists are often concerned with understanding the different microscopic agents of these systems so that they may explain or predict future macroscopic behaviors. Non-equilibrium statistical physics turns towards models and simulations to capture macroscopic system behavior accounting for varying amounts of microscopic contributors.

This project analyses different stochastic non-equilibrium models of transcription ranging in complexity from a simple birth and death model to a multi-state model. First, a biological background of transcription gives the reader necessary information to understand the structure of the following models. Next each model has its applications discussed, is outlined mathematically, and is analyzed for different probabilistic rates. The order of the thesis is such: the one state model (Chapter 3), the two state model (Chapter 4), and the three state model (Chapter 5). Other possible models and future goals of this project are discussed at the end of the paper (Chapters 5 and 6).

Chapter 2

Biological Background and the Physics Connection

2.1 mRNA Transcription

Physicists in the modern area are interested in how interactions of microscopic particles influence their macroscopic behavior. An analogous point of interest for Biologists is the relationship between intercellular functions and their macroscopic presentations in tissues, organs, and organ systems. One cellular process that is at the heart of cell operations is deoxyribonucleic acid (DNA) transcription. To generalize, transcription is the first step towards protein creation utilizing cell DNA. It is important to understand that protein creation then determines cell composition and functioning; any process that guides protein creation, both what proteins are made and how much of any one protein is produced, therefore has a direct impact on a cellular processes[9].

DNA is a polymer, consisting of "nucleotide" units which are comprised of a sugar, a nitrogen-containing base, and a phosphate group. DNA is a double helix consisting of two strands running antiparallel to one another. Adjacent nucleotides on a single DNA strand share a strong bond creating a structurally sound backbone; nucleotides parallel to one another on opposite strands share a weaker chemical bond, two or three hydrogen bonds, allowing the two strands to separate when needed, an example of which is transcription. Figure 2.1 displays a visual of transcription. The 3 prime (3') and 5 prime (5') denote directionality of the DNA stands based upon the orientation of the nucleotide bases; figure 1 visualizes the strands antiparallel orientation. The nitrogen-containing bases, a nitrogenous base, can have one of four bases: adenine (A), guanine (G), cytosine (C), or thymine (T) in DNA. Base G pairs with base C on opposite strands whereas base A pairs with base T[9] [10].

Ribonucleic acid (RNA) is another key component in transcription. RNA is responsible for the facilitation of transcription and is a product of transcription. RNA is similar to DNA. However, key differences include the single-stranded nature of RNA and that RNA has the nitrogenous base uracil (U) in substitution of base T in DNA. As visualized in Figure 2.1, the protein RNA polymerase, unwinds the DNA, and builds a chain of pre-messenger RNA (pre-mRNA) out of ribonucleotide triphosphates. Messenger RNA (mRNA) is the result of the processing of pre-mRNA; it is allowed outside of the cell nucleus in a eukaryote. Eukaryotes are cells that have a nucleus

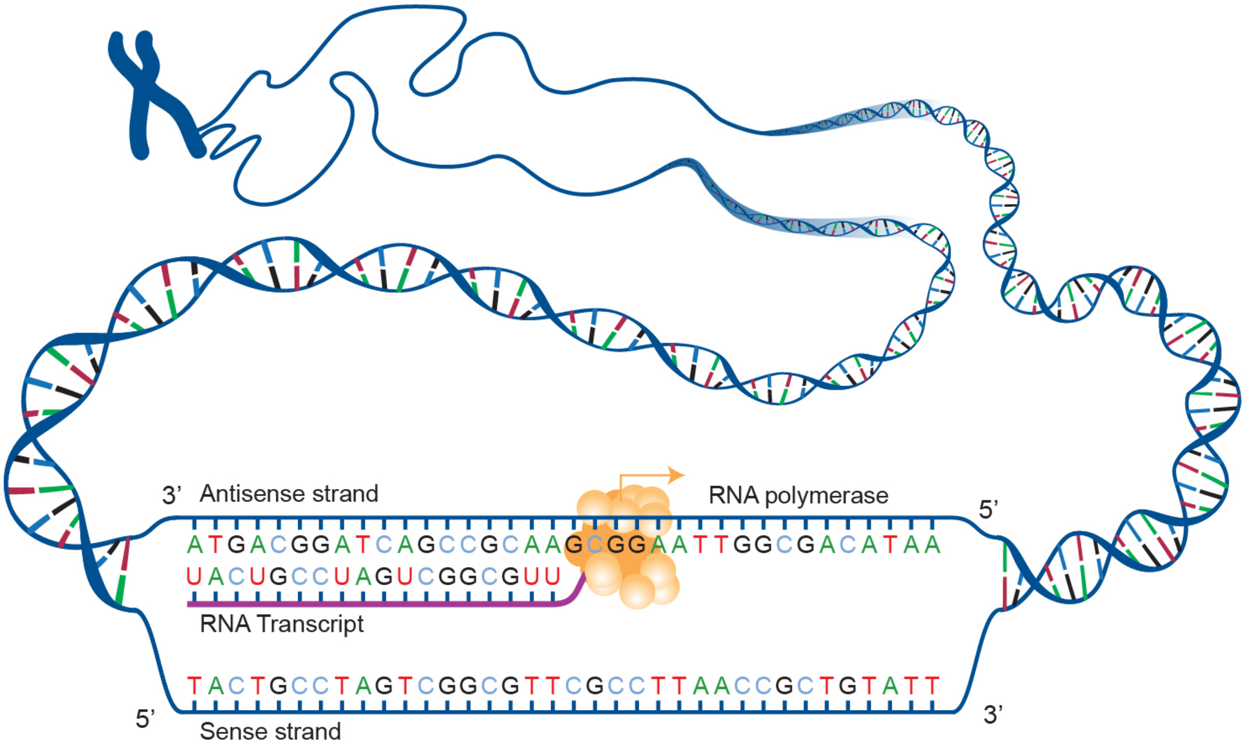


Figure 2.1: Labelled diagram of a DNA strand undergoing transcription[10].

housing DNA rather than it free floated in intracellular space as it does in prokaryotes. The mRNA processing that occurs is three-fold: there is a 5’ cap and 3’ tail added to prevent degradation and there are identified unnecessary portions of the pre-mRNA that are removed. The biological process of transcription is then complete and a protein is ready to be created through translation[9][10].

It is important to note that this process, both the initiation of transcription and processing of mRNA, play a huge role in gene regulation. Understanding when and how much of a protein is made can shine light on the nature of micro-anatomical physiology. On a cellular level these proteins keep the gene alive. On a tissue, organ, and organism level these proteins differentiate cell lines and keep the complex machine, like the human body, functioning. Errors in transcription regulation can lead to an access of protein, or a depletion, which can throw an organism out of equilibrium and into disease[9][10]. A well-known example of this is tumerogensis, which can develop into cancerous legions, the over-expression of tumor-causing genes[8].

Chapter 3

One-State Model for mRNA Transcription

Statistical physics studies system behavior comprised of a large number of agents. A mere 18g of H_2O consists of 6.023×10^{23} molecules of water, which is the well known and widely applied Avogadro number; the human genome is comprised of billions of base pairs. Biological systems, from the micro-anatomical to population studies, lend themselves to exploration via statistical physics. Just as a cell acts differently than an organ, and a individual separately from its community, individual components alone do not represent the behavior of the collective[11]. Statistical physics works to model and predict system behavior based on the input of its microscopic constituents interacting among themselves and with their environment. This is not an easy task to undergo due to the near endless possibilities that exist for each component. Imagine a game of musical chairs: there are a definite number of chairs and children yet a multiplicative amount of assembly possible when the music stops and children rush to find an open space. Increase this toy model to one with scales of 10^{23} and higher. This example elucidates the reasons why many values sought after in statistical physics involve total averages and probabilities which help contextualize the potential outcomes[11].

Systems move toward thermal equilibrium, their most energy efficient form, if given the proper time. A common instance of such would be reaching for a cup of scalding coffee to find instead a drink that has cooled to room temperature. Credit for the theoretical framework of predicting the behavior of a system in equilibrium goes to late-ninetieth physicists J.W. Gibbs and Ludwig Boltzmann[3]. In the 1970s such methodology was refined and today equilibrium statistical physics finds itself as an integral part to any college physics curriculum[11].

Alas, life outside a lab does not maintain such well-behaved equilibrium with system properties in flux with exposure to a changing environment. Exchange of matter, energy, and/or information leads to non-equilibrium behavior. Arising from inquiry and pursuit of understanding, the field of non-equilibrium physics studies the time evolution of complex systems well outside thermal equilibrium.

A *master equation*—which describes the dynamics of a system—can be written and sometimes solved provided there is a valid model for the system where interactions among the agents and evolution between states are known. Solving these equations can be challenging, or not feasible at times, yet solutions for simple models can provide insight into evolving patterns and fundamental features of the non-equilibrium behaviors of these systems.[11] If interested in a deeper survey of the kinetic approach to non-equilibrium systems, please refer to Kraphivsky, Redner and Ben-Naim's book *"A kinetic view of statistical physics"* [7].

Many processes within fields such as physics, chemistry, and biology are stochastic(probabilistic). In fact, it is uncommon to find a system that is completely deterministic. *Birth-and-death processes*—also referred to as *generation-recombination* or *one-step processes*—are important for modeling systems such as photon emission/absorption, chemical reactions and population dynamics[7].

Among the best known *one-step processes* are *random walks* illustrated in Figure 3.1. Typically they include an agent on a lattice that moves randomly to an allowed lattice space specified by model constraints. This is a well studied problem yet can be applied to new puzzles and be applied outside of traditional physics study.

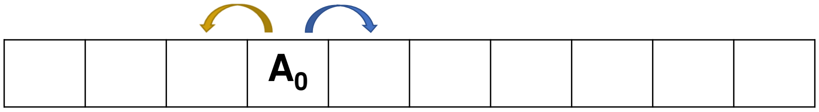


Figure 3.1: Diagram of a *random walk* model. Agent A_0 can move either to the left or right, denoted by the corresponding gold and blue arrows, and will randomly move in one direction. The next iteration of the model A_0 will undergo another random movement along the lattice to one of the adjacent spaces.

"Birth and death processes" can be studied analytically using a master equation. The master equation is a balance (continuity) equation which expresses the conservation of configuration probabilities. A *configuration* is a snapshot of the system. The master equation describes the evolution of all possible configurations into other configurations, as well as their origins. In general, consider a system in state “r” at time “t”. $P_r(t)$ is the probability that this system is in this particular state at time “t”. The time dependence of P_r is given by the master equation:

$$\frac{dP_r}{dt} = \sum_s P_s W_{sr} - \sum_s P_r W_{rs} \tag{3.1}$$

The probability of state “r” increases with time due to all states that evolve into state “r”, and it decreases with time because of transitions from state “r” to other states. In this equation, W are the transition rates to and from state “r”. Knowing W allows us to calculate all probabilities P_r as a function of time.

The transition rates are customized depending on the type of physical system that is being studied. Let us illustrate this equation with the simple example of a symmetric random walk on a one-dimensional lattice. The walk is described by the probability $P_n(t)$ that the walk is at site n at time t . The probabilities of moving left and right are the same and equal to 1. This probability evolves as:

$$\frac{dP_n}{dt} = P_{n+1} + P_{n-1} - 2P_n \tag{3.2}$$

The first two terms on the right hand side account for the increase in probability P_n because of a hop from $n-1$ to n or because of a hop from $n+1$ to n, respectively. Similarly, the last term accounts for the decrease of P_n because of hopping from n to $n \pm 1$.

For the generic ”birth and death process” represented in Figure.1, the associated master equation is:

$$\begin{aligned} \frac{dP_0}{dt} &= -\lambda_0 P_0(t) + \mu_1 P_1(t) \qquad (3.3) \\ \frac{dP_n}{dt} &= \lambda_{n-1} P_{n-1}(t) - (\lambda_n + \mu_n) P_n(t) + \mu_{n+1} P_{n+1}(t) \qquad (3.4) \end{aligned}$$

Here λ_0 is the rate at which the system evolves from P_0 to P_1 and μ_1 is the rate the system moves from P_1 to P_0 and so forth for increasing amounts of P_n. The rates λ_n and μ_n are called the birth rates and death rates respectively (n is the population size), and are positive numbers. For example, in a growing population, once the number n of individuals is zero, the growth process stops . Thus the state $n = 0$ is a special state called an ”absorbing state”.

The master equation is not easily solvable, but for special cases there are analytical solutions. The simplest case is a pure birth process or the ”Poisson process”. In this case, we have all $\lambda_n = \lambda$, and all $\mu_n = 0$. The solution for probability P_n is the Poisson distribution:

$$P_n = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \tag{3.5}$$

This analytical result can be found using the generating function technique outlined in [7].

3.1 Definition of mRNA One-State Model

3.1.1 Deterministic One-State Model

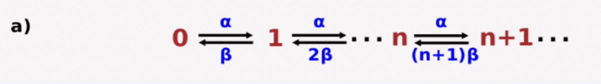


Figure 3.2: A sketch of a deterministic one-state model. Image adapted from Klindziuk's "Stochastic Modeling of DNA Transcription and Gene Expression"[6].

A very basic deterministic one-state model is described by the following equation, and depicted in Figure 3.2:

$$\frac{dn}{dt} = \alpha - \beta n \tag{3.6}$$

where n is the number of mRNAs that are being created with a "birth" rate α and lost with a "death" rate β. This is a first-order approximation of a stochastic process, which is a preferred model for capturing the biochemical processes of mRNA transcription[12]. The death rate would symbolize two different methods of mRNA's "dying". One would be mRNA degradation which is a cellular process that results in the digestion of mRNAs; the second would be finalized mRNA's leaving the nucleus of a cell to be translated if measuring mRNA products in a eukaryotic cell.

This equation has a simple analytical solution dependent on the two rates and the initial number of mRNAs, n_0:

$$n(t) = \frac{\alpha}{\beta}(1 - e^{-\beta t}) + n_0 e^{-\beta t} \tag{3.7}$$

Examples of different cases are depicted in Figures 3.4 and 3.5. Figure 3.4 shows the outcome of a large birthrate and minimal death rate. The function decreases exponentially, as it does in Figure 3.5 where alpha and beta are equal, before reaching a steady state value. That steady state value is discussed in further detail below.

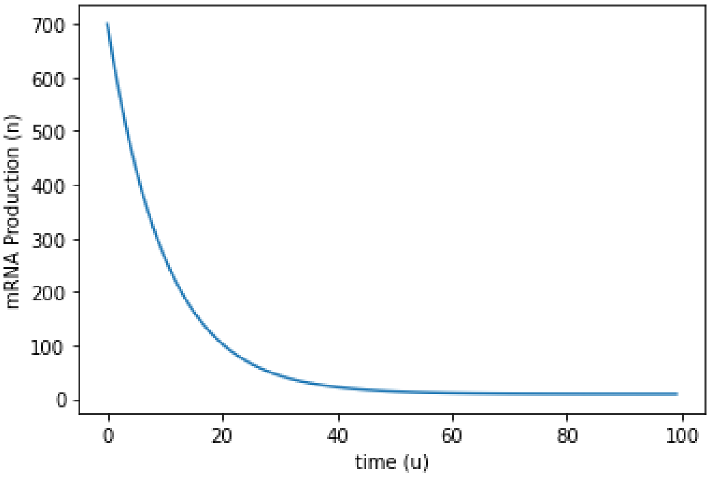


Figure 3.3: Total number of mRNAs , n, as a function of time (in arbitrary units) with corresponding birth and death rates of $\alpha = 0.9$ and $\beta = 0.1$. Initial number of mRNAs present in population is 700.

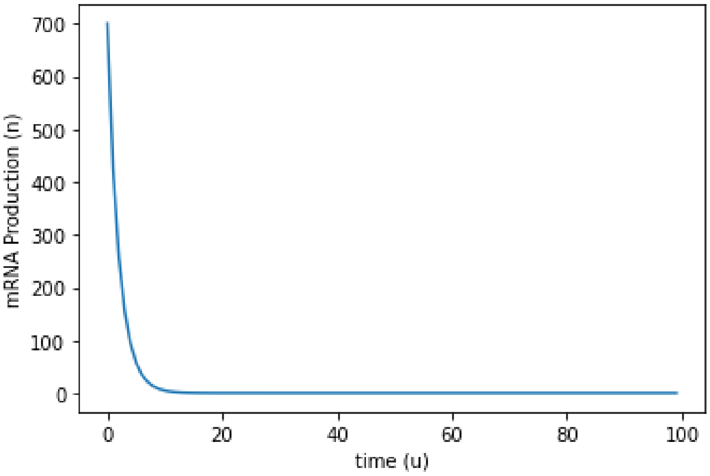


Figure 3.4: Total number of mRNAs , n, as a function of time (in arbitrary units) with corresponding birth and death rates of $\alpha = 0.5$ and $\beta = 0.5$. Initial number of mRNAs present in population is 700.

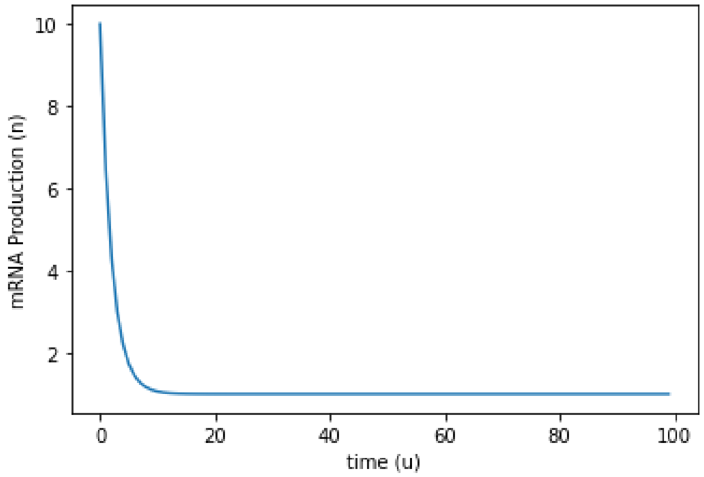


Figure 3.5: Total number of mRNAs , n, as a function of time (in arbitrary units) with corresponding birth and death rates of $\alpha = 0.5$ and $\beta = 0.5$. Initial number of mRNAs present in population is 10.

A common interest in statistical physics is studying the behavior of a system at specific limiting cases. An example here would be the behavior of a system as t approaches infinity. The result is shown below utilizing our master equation.

$$n(t)_{t \to \infty} = \frac{\alpha}{\beta}(1 - e^{-\beta t}) + n_0 e^{-\beta t} = \frac{\alpha}{\beta} \tag{3.8}$$

Here it is apparent that long-term behavior of transcription for the one state model is dependent on the relationship between the birth and death rate. This is highlighted by Figure 3.5, Figure 3.6, and equation 3.8. For these figures α and β are the same which leads to a steady state production of one mRNA present as the time duration grows. In Figure 3.5 this is not as apparent but taking a closer look, when the initial number of mRNA's present is lower, Figure 3.6 displays this overall trend showing the function decreases exponentially to one when reaching steady state.

3.1.2 Stochastic One-State Model

The stochastic one-state model for mRNA production is a special case of a general "birth and death" process presented above, with the associated master equation:

$$\frac{dP_0}{dt} = -\lambda_0 P_0(t) + \mu_1 P_1(t) \tag{3.9}$$

$$\frac{dP_n}{dt} = \lambda_{n-1} P_{n-1}(t) - (\lambda_n + \mu_n) P_n(t) + \mu_{n+1} P_{n+1}(t) \tag{3.10}$$

For the one-state model, this master equation reduces to:

$$\begin{aligned} \frac{dP_0}{dt} &= \beta P_1(t) && (3.11) \\ \frac{dP_n}{dt} &= -\alpha P_n(t) + \beta(n+1)P_{n+1}(t) && (3.12) \end{aligned}$$

$P_n(t)$ represents the probability of having n mRNA transcripts at time t. This master equation is similar to a random walk problem. The steady-state probability distribution of the number of produced mRNA molecules is given by the Poisson solution derived using the generating-function technique:

$$P_n = \frac{(\frac{\alpha}{\beta})^n}{n!} e^{-\frac{\alpha}{\beta}} \quad (3.13)$$

Knowing this probability distribution, we can now calculate the average number of mRNA molecules:

$$< n >= \sum n P_n = \frac{\alpha}{\beta} \quad (3.14)$$

The average number of mRNA transcripts for the steady state matches the result for the deterministic model for t going to infinity. Figure 3.7 shows the multitude of steady state values for differing rates of alpha and beta. Figure 3.7 illustrates that higher birth rates and lower death rates result in the highest steady state values. A future improvement of this model would be to shift away from the probabilistic rates utilized here, with α and β ranging from zero to one, to biologically accurate rates that have units that are not arbitrary.

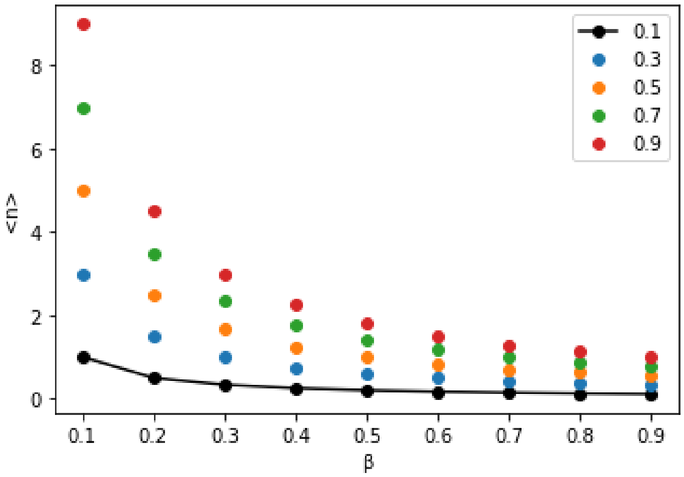


Figure 3.6: Graphing of steady state for mRNA production vs β dependent on α value. Colors correspond to differing values of α; the black line is used for visual reference and trend identification rather than having mathematical significance.

Chapter 4

Two-State Models for mRNA Transcription

4.1 Two- State Models in Physics

While the simple one state model is useful to consider very basic organism's transcriptions activity, the majority of transcription occurs at a increasingly more nuanced level. A *two state process*—also termed a *random telegraph model* or a *two-state Markov chain*—can account for two different action states of a system: one where the system is ON and one where it is OFF. An interesting analogy for this type of model is the quantum state of a particle. Measurement of a particle can change its quantum state and governing equations in the field can give probabilistic predictions of what state a particle will be in upon observation.

4.2 Definition of mRNA Two-State Model

This model of transcription predicts overall mRNA production and transcription activity when temporally their are periods of active and non-active work done by the DNA polymerase. Homeostasis of temperature, energy storage, and protein composition within cells are all integral to maintaining life; homeostasis is maintained by the regulation of genes and the ability of the organism to respond to its environment appropriately.

This two-state model mirrors real gene regulation for genes that have promoters. Promoters are sequences that come before the coding region of a gene; for transcription to begin there must be the proper factors present at the site triggering initiation of transcription. In real life, this form of turning genes "ON" or "OFF" is highly dependent on the interaction of the organism with the surrounding environment.

The two-state model outlined below was introduced by J. Peccoud [?] in 1995, as a Markovian birth and death model of gene product synthesis [?]. This study is considered a seminal work in the field of gene expression modeling, as it generated a plethora of studies. This simple model considers each gene in two states: Active (A) and Inactive. The switch between the states happens based on the following rules:

- $I \to A$ with rate λ
- $A \to I$ with rate μ
- $A \to A + N$ with rate ν
- $N \to \emptyset$ with rate δ

N is the mRNA being generated, and $\emptyset$ is the zero of the chemical reaction, meaning that the protein is transformed into something else that is not considered in the model.

We now define the following quantities:

- In any state (i, n) the first coordinate denotes the status of the gene, with $i = 0$ when the gene is inactive, and $i = 1$ when the gene is active. The second coordinate n represents the number of molecules (copies) of the mRNA N in the cell.
- $P_{0,n}$ is the probability that at time t the gene is inactive and that n molecules of N are present.
- $P_{1,n}$ is the probability that at time t the gene is active and that n molecules of N are present.

The associated master equation can be split for two cases, as presented in [13]:

For $\forall n \geq 0$:

$$\begin{aligned} \frac{dP_{0,n}}{dt} &= -(\lambda + n\delta)P_{0,n}(t) + (n+1)\delta P_{0,n+1}(t) + \mu P_{1,n}(t) && (4.1) \\ \frac{dP_{1,0}}{dt} &= -(\mu + \nu)P_{1,0}(t) + \delta P_{1,1}(t) + \lambda P_{0,0}(t) && (4.2) \end{aligned}$$

For $\forall n \geq 1$:

$$\frac{dP_{1,n}}{dt} = -(\mu + \nu + n\delta)P_{1,n}(t) + (n+1)\delta P_{1,n+1}(t) + \nu P_{1,n-1}(t) + \lambda P_{0,n}(t) \quad (4.3)$$

To exemplify how this works, let's pick first $n = 0$, which means no copies of mRNA are produced. The system of equations then becomes:

$$\begin{aligned} \frac{dP_{0,0}}{dt} &= -\lambda P_{0,0}(t) + \delta P_{0,1}(t) + \mu P_{1,0}(t) && (4.4) \\ \frac{dP_{1,0}}{dt} &= -(\mu + \nu)P_{1,0}(t) + \delta P_{1,1}(t) + \lambda P_{0,0}(t) && (4.5) \\ & && (4.6) \end{aligned}$$

We now pick $n = 1$, which means only one molecule (copy) of mRNA is produced. The system of equations then becomes:

$$\begin{aligned} \frac{dP_{0,1}}{dt} &= -(\lambda + \delta)P_{0,1}(t) + 2\delta P_{0,2}(t) + \mu P_{1,1}(t) && (4.7) \\ \frac{dP_{1,0}}{dt} &= -(\mu + \nu)P_{1,0}(t) + \delta P_{1,1}(t) + \lambda P_{0,0}(t) && (4.8) \\ \frac{dP_{1,1}}{dt} &= -(\mu + \nu + \delta)P_{1,1}(t) + 2\delta P_{1,2}(t) + \nu P_{1,0}(t) + \lambda P_{0,1}(t) && (4.9) \end{aligned}$$

4.2.1 No protein degradation, $\delta = 0$

In this case, the master equation simplifies to the following:

For $n = 0$:

$$\frac{dP_{0,0}}{dt} = -\lambda P_{0,0}(t) + \mu P_{1,0}(t) \tag{4.10}$$

$$\frac{dP_{1,0}}{dt} = -(\mu + \nu) P_{1,0}(t) + \lambda P_{0,0}(t) \tag{4.11}$$

$$\tag{4.12}$$

For $n = 1$:

$$\frac{dP_{0,1}}{dt} = -\lambda P_{0,1}(t) + \mu P_{1,1}(t) \tag{4.13}$$

$$\frac{dP_{1,0}}{dt} = -(\mu + \nu) P_{1,0}(t) + \lambda P_{0,0}(t) \tag{4.14}$$

$$\frac{dP_{1,1}}{dt} = -(\mu + \nu) P_{1,1}(t) + \nu P_{1,0}(t) + \lambda P_{0,1}(t) \tag{4.15}$$

If we put them together into one system of differential equations:

$$\frac{dP_{0,0}}{dt} = -\lambda P_{0,0}(t) + \mu P_{1,0}(t) \tag{4.16}$$

$$\frac{dP_{0,1}}{dt} = -\lambda P_{0,1}(t) + \mu P_{1,1}(t) \tag{4.17}$$

$$\frac{dP_{1,0}}{dt} = -(\mu + \nu) P_{1,0}(t) + \lambda P_{0,0}(t) \tag{4.18}$$

$$\frac{dP_{1,1}}{dt} = -(\mu + \nu) P_{1,1}(t) + \nu P_{1,0}(t) + \lambda P_{0,1}(t) \tag{4.19}$$

We can solve the system of equations using the Python ODE solver.

Two different sets of parameters are chosen for graphical representation. Figure 4.1 displays probabilities dependent on rates that would result in low amounts of gene activation and mRNA production. As expected, the probability of being is state P_{00} is maintained across time and has values than any other state. Figure 4.2 on the other hand, visualizes the probabilities of different states when the rate of gene activation and mRNA production is high. P_{10} and P_{11} are both initially high in value and P_{00} exponentially declines. The peaks observed in Figure 4.2 represent a higher probability of being in some state P with one mRNA molecule present, either P_{01} or P_{11}. This peak phenomena can be attributed to the fact that birth rate ν and activation rate λ are high whereas degradation rate μ is low in Figure 4.2.

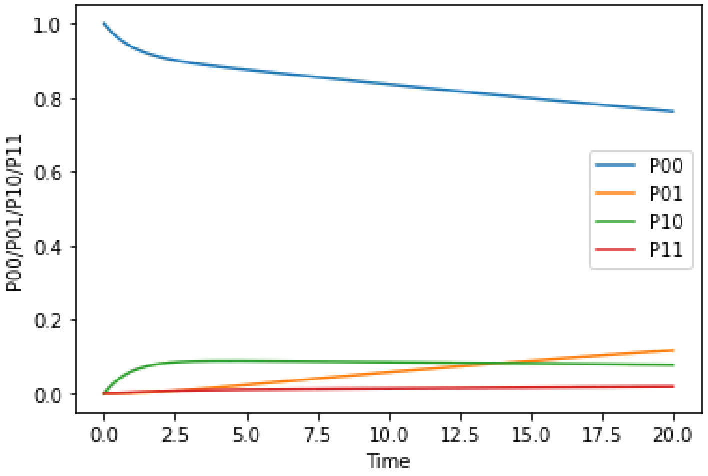


Figure 4.1: Displays two-state probabilities, with corresponding birth and death rates of $\lambda = 0.1$, $\mu = 0.9$, and $\nu = 0.1$.

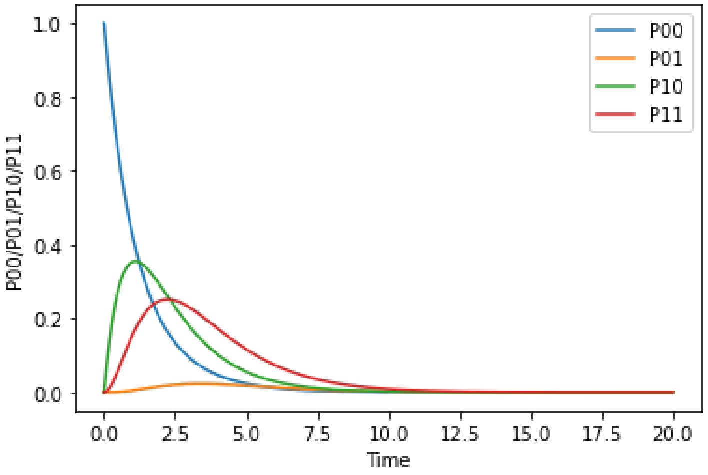


Figure 4.2: Displays two-state probabilities, with corresponding birth and death rates of $\lambda = 0.9$, $\mu = 0.1$, and $\nu = 0.9$.

4.2.2 Protein degradation, $\delta \geq 0$

In this case the previous system of equations for $n = 0, 1$ becomes:

$$
\begin{aligned}
\frac{dP_{0,0}}{dt} &= -\lambda P_{0,0}(t) + \delta P_{0,1}(t) + \mu P_{1,0}(t) && (4.20)\\
\frac{dP_{0,1}}{dt} &= -(\lambda + \delta) P_{0,1}(t) + 2\delta P_{0,2}(t) + \mu P_{1,1}(t) && (4.21)\\
\frac{dP_{1,0}}{dt} &= -(\mu + \nu) P_{1,0}(t) + \delta P_{1,1}(t) + \lambda P_{0,0}(t) && (4.22)\\
\frac{dP_{1,1}}{dt} &= -(\mu + \nu + \delta) P_{1,1}(t) + 2\delta P_{1,2}(t) + \nu P_{1,0}(t) + \lambda P_{0,1}(t) && (4.23)
\end{aligned}
$$

If we assume $\delta > 0$, in order to be able to close the system of equations we will need to make the assumption that the probabilities $P_{0,2}$ and $P_{1,2}$ are 0, meaning that there is no more than one copy of mRNA in the system. So the system of equations simplifies to:

$$
\begin{aligned}
\frac{dP_{0,0}}{dt} &= -\lambda P_{0,0}(t) + \delta P_{0,1}(t) + \mu P_{1,0}(t) && (4.24)\\
\frac{dP_{0,1}}{dt} &= -(\lambda + \delta) P_{0,1}(t) + \mu P_{1,1}(t) && (4.25)\\
\frac{dP_{1,0}}{dt} &= -(\mu + \nu) P_{1,0}(t) + \delta P_{1,1}(t) + \lambda P_{0,0}(t) && (4.26)\\
\frac{dP_{1,1}}{dt} &= -(\mu + \nu + \delta) P_{1,1}(t) + \nu P_{1,0}(t) + \lambda P_{0,1}(t) && (4.27)
\end{aligned}
$$

Figure 4.3 and 4.4 demonstrate complicated behavior within the two-state model. Figure 4.3 has moderate values for all rates with lower mRNA production. It is interesting to note, that P_{00} and P_{10} maintain higher values than either of the probabilities of being in a state with one mRNA. Even when further increasing the rate μ and decreasing mRNA produced, ν, probability P_{10} and P_{11} show initial peaks and significant long term behavior. Further exploration of the complex interaction between these rate based states could prove enlightening.

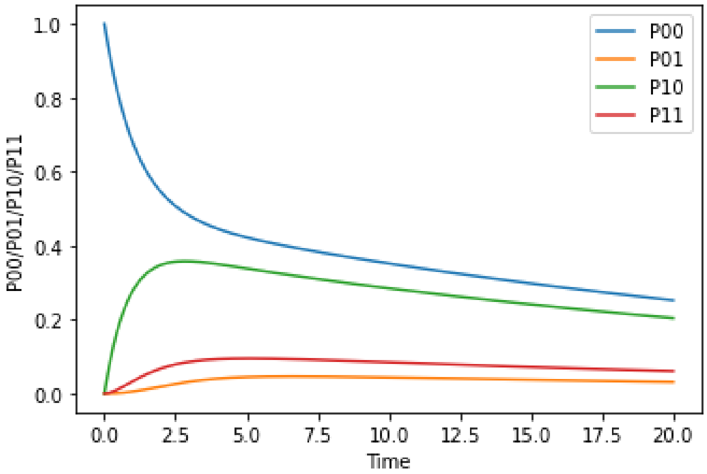


Figure 4.3: Displays two-state probabilities, with corresponding birth and death rates of $\lambda = 0.5$, $\delta = 0.5$, $\mu = 0.5$, and $\nu = 0.3$.

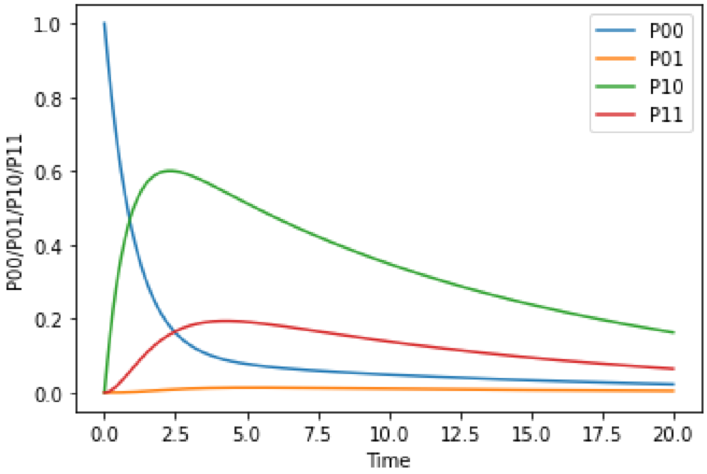


Figure 4.4: Displays two-state probabilities of switching between active and inactive as well as total mRNA production. Corresponding birth and death rates are $\lambda = 0.5$, $\delta = 0.5$, $\mu = 0.9$, and $\nu = 0.1$.

Figures 4.5 and 4.6 depict the results of the Monte Carlo simulations for the two-state model. The code outlining the Monte Carlo simulation for the two-state model can be found in the Appendix Figure 8.2.. The figures display three rates: n_A, whether or not the gene is active; n_I, whether or not the gene is inactive; and n_P the total number of proteins produced assuming every one mRNA creates one protein. With this parallel, n_P could also represent the total number of mRNA present at any given time.

Figure 4.5 displays a clear correlation between longer periods of n_A valued at 1, the gene is active, and corresponding mRNA production. The mutli-state model will explore this relationship further and will connect to the probabilistic plots generated from the ODE solver method as well.

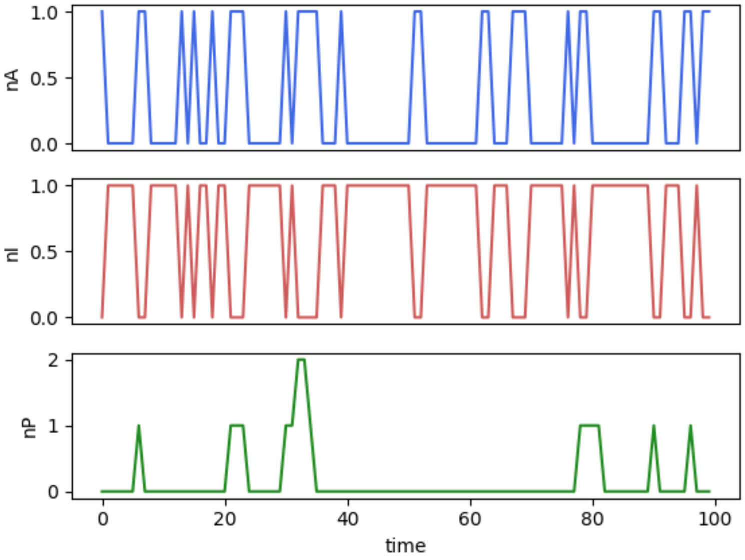


Figure 4.5: Monte Carlo simulation results for one gene that switches back and forth from inactive to active states and produces mRNA. Corresponding birth and death rates are $\lambda = 0.5$, $\delta = 0.5$, $\mu = 0.5$, and $\nu = 0.5$.

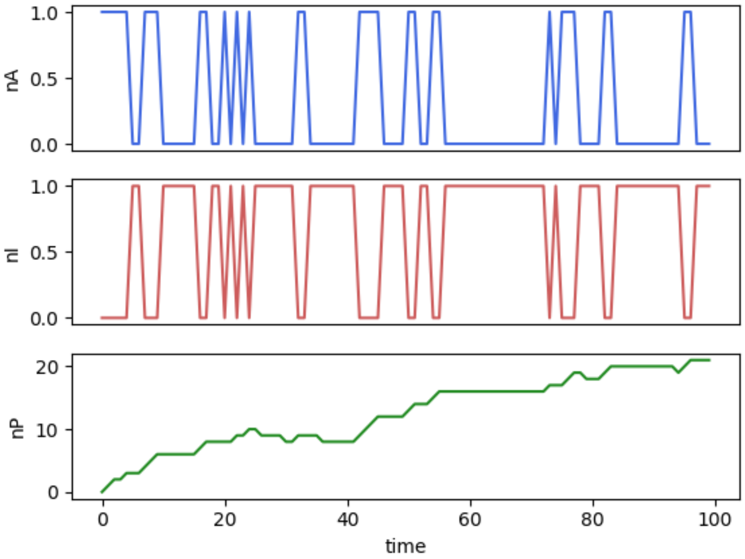


Figure 4.6: Monte Carlo simulation results for one gene that switches back and forth from inactive to active states and produces mRNA. Corresponding birth and death rates of $\lambda = 0.5$, $\delta = 0.5$, $\mu = 0.9$, and $\nu = 0.1$.

4.3 Definition of ”Poisson with Zero Spike” Model

The Poisson with Zero Spike model is a variation of the two state model by which the ”Off” state now can only have zero mRNA molecules. The ”On” state can have as many mRNA molecules as the one state model have have multiple configurations[6]. This model would be applicable to situations where the probability of switching between the states is low. Further analysis could compare low rates of λ and μ in the two state model to the Poisson with Zero Spike model presented here. Below outlines the mean number of mRNAs produced where γ represents a transition rate between states and x is equal to α over β.

$$< n >= \frac{xe^x}{(e^x + \gamma)} \tag{4.28}$$

This model could represent a constitutive gene, one continuously expressed. These genes are vital for cell maintenance and function in all conditions and would have very low rates of switching to the ”Off” state.

Chapter 5

Three-State Model for mRNA trascription

5.1 Three- State Models in Physics

As we have seen in the previous chapters, two-state models proved to be extremely useful in studying a variety of physical systems, from magnetism to biophysics or social systems. But sometimes the two-state models do not capture the essential features of the physical systems, and more general models are necessary.

For example, a generalization of the Ising model is the Potts model—fully named the *q-state Potts model*. This model is beloved in the field of equilibrium statistical physics for modeling more complex systems. It dates back to two mathematicians in the 1900s, Julius Ashkin and Edward Teller and then expanded to what it is today by Renfrey B. Potts. The Potts model is primed to examine different spin configurations of a larger lattice, where spins can have three orientations, or states. This maintains the basis for any good statistical physics model, as the model allows for the study of macroscopic behaviors from studying the microscopic internal elements[1].

The orientations of the spins along the lattice can represent, in a different context, a variety of other features of the system that is being considered. For example, for a voter model, a spin up can be a vote ”yes”, a spin down is a vote ”no” and horizontal spin can be associated with a neutral state or an ”abstention”. Drawing ties to this thesis, spins could be assigned ”on” or ”off” to represent the activation of transcription within a gene, such as one ”on” state and two ”off” states, as presented in [2].

The Potts model can be used to study a wide variety of system behaviors. As mentioned before one application could be to study epidemics and population health, but it can also be applied as a model for tumor growth. Beaudin in [1] also outlines an example where the Potts model is used to study human interaction tying the model to sociology studies. The cellular Potts model has been extensively used to model interacting cell systems at the tissue-level. [4]. The Potts model introduces our discussion of the three state model applied to transcription regulation and the ways multi-state models can more precisely emulate the cellular mechanics observed.

5.2 Definition of mRNA Three-State Model

Physicists often seek balance between making models accurate and making models solvable when studying the world around them. The *three-state model* of mRNA production is no different. While analysing the activation of initiation of a gene, as either on or off, the nuance of gene regulation is simplified to a high degree. This model introduces another factor, another state, at which transcription is turned OFF. Interplay between the three states then allows genes with higher levels of environmental regulation to be modeled. Specifically, when introducing this third option as one that turns OFF a gene, the model becomes best suited towards describing a repressor; binding near the site of transcription, repressors are molecular structures that halt transcription. These repressors can be introduced to the cellular environment from external signaling or result from a down stream affect of mRNA production forming a negative feedback loop. Both cases will result in transcription halting and mRNA production adjusting to environmental cues to maintain homeostasis for proper cellular functioning.

Another variant of the three state model exists when adding an addition state that can occur once the gene is active, or ON. This would describe the activity of an enhancer, which acts in contrast to a repressor by motivating increased transcriptional activity. To remain viable in fluctuating environments, cellular organisms must be able to adapt. As genes control proteins, which then control nearly all cellular function, being able to modify transcription of genes is the first step towards behavioral modification and environmental response.

We focus in this chapter on the three-state model presented in [2] and represented in the following diagram:

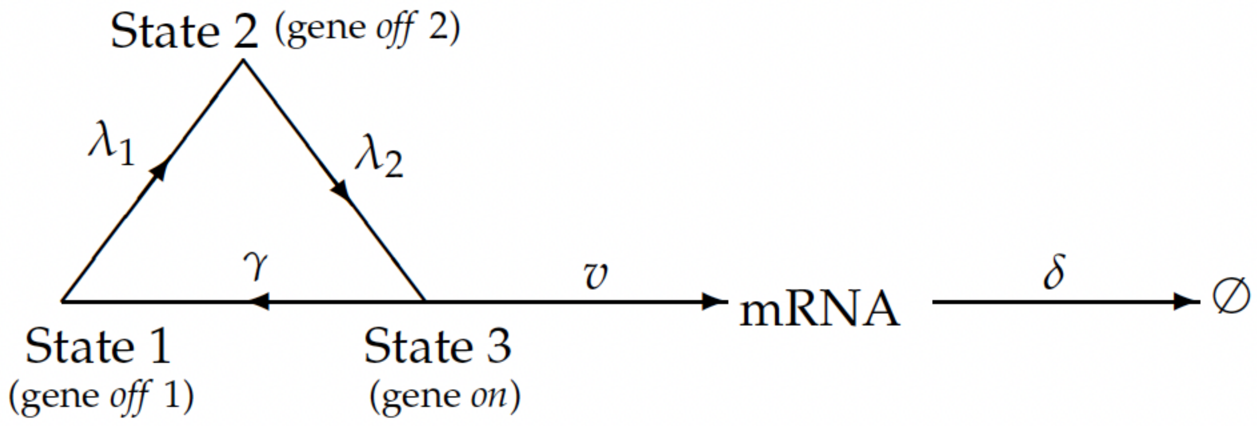


Figure 5.1: Diagram of the 3-state model. From "A Novel Approach for Calculating Exact Forms of mRNA Distribution in Single-Cell Measurements" by Chen and Jaio [2].

As presented in [2], the gene can be in two "off" states and one "on" state. The transition between the states happen with the probabilities λ_1, λ_2 and γ. One the gene is in State 3, the "on" state, copies of mRNA are being generated, which eventually degrade into proteins.

The differential equations that govern this system are written for the probabilities that the gene resides in state $i = 1, 2, 3$ with m mRNA copies being produced at time t:

$$\frac{dP_{m,1}}{dt} = \gamma P_{m,3} - (m\delta + \lambda_1)P_{m,1}(t) + (m+1)\delta P_{m+1,1}(t) \tag{5.1}$$

$$\frac{dP_{m,2}}{dt} = \lambda_1 P_{m,1} - (m\delta + \lambda_2)P_{m,2}(t) + (m+1)\delta P_{m+1,2}(t) \tag{5.2}$$

$$\frac{dP_{m,3}}{dt} = \lambda_2 P_{m,2} - (m\delta + \gamma + v)P_{m,3}(t) + (m+1)\delta P_{m+1,3}(t) + vP_{m-1,3}(t) \tag{5.3}$$

$$\tag{5.4}$$

For initial conditions, we assume $P_{0,1}(0) = 1$, $P_{0,2}(0) = 0$, $P_{0,3}(0) = 0$ and all the other probabilities $P_{m,i}(0) = 0$.

5.3 Computer Simulations

We study this model two different ways, via Monte Carlo simulations and via solving numerically using the Python ODEINT module the system of differential equations for specific cases. The code outlining the Monte Carlo simulation for the three-state model can be found in the Appendix Figure 8.2.

For the ODE solver method, the system of equations presented above is customized for the specific number of mRNA copies that is being considered. For example, for $m = 1$ copies, we will need a system of 6 differential equations such as:

$$\frac{dP_{0,1}}{dt} = \gamma P_{0,3} - (m\delta + \lambda_1)P_{0,1}(t) + \delta P_{1,1}(t) \tag{5.5}$$

$$\frac{dP_{0,2}}{dt} = \lambda_1 P_{0,1} - \lambda_2 P_{0,2}(t) + \delta P_{1,2}(t) \tag{5.6}$$

$$\frac{dP_{0,3}}{dt} = \lambda_2 P_{0,2} - (\gamma + v)P_{0,3}(t) + \delta P_{1,3}(t) \tag{5.7}$$

$$\frac{dP_{1,1}}{dt} = \gamma P_{1,3} - (\delta + \lambda_1)P_{1,1}(t) \tag{5.8}$$

$$\frac{dP_{1,2}}{dt} = \lambda_1 P_{1,1} - (\delta + \lambda_2)P_{1,2}(t) \tag{5.9}$$

$$\frac{dP_{1,3}}{dt} = \lambda_2 P_{1,2} - (\delta + \gamma + v)P_{1,3}(t) + vP_{0,3}(t) \tag{5.10}$$

We assume that all the probabilities of having more than one mRNA copies are zero.

Figures 5.2 and 5.4 both display resulting gene activity when rates of activation are high but mRNA production is low. Figure 5.2 demonstrates this with P_{03} maintaining the highest probability for the majority of the time assessed. This is further illustrated in Figure 5.4 by the low amounts of mRNA produced. When elevating the rates of mRNA production, Figure 5.3 shows the increase probabilities of state P_{11}, P_{12}, and P_{13}. Figure 5.5 demonstrates mRNA production for a case of more constant growth, with little degradation, and shows a correlation between mRNA production plateaus and periods of increased inactivation.

Figure 5.4 displays staggered mRNA production with peaks appearing after longer durations of n_{l2} and n_A. Overall the number of mRNA produced is low as rates γ, λ_1, and δ counteract the higher valued rates of $lambda_2$ and ν. Figure 5.5 shows Monte Carlo simulation results of a steady production of mRNA production. Figure 5.5 has the exact same rates as Figure 5.4 except that here $\delta = 0$. The impact of removing mRNA degradation is strongly capture in the Monte Carlo simulation results of Figure 5.5 with total mRNA product reaching levels ten times the amount seen in Figure 5.4.

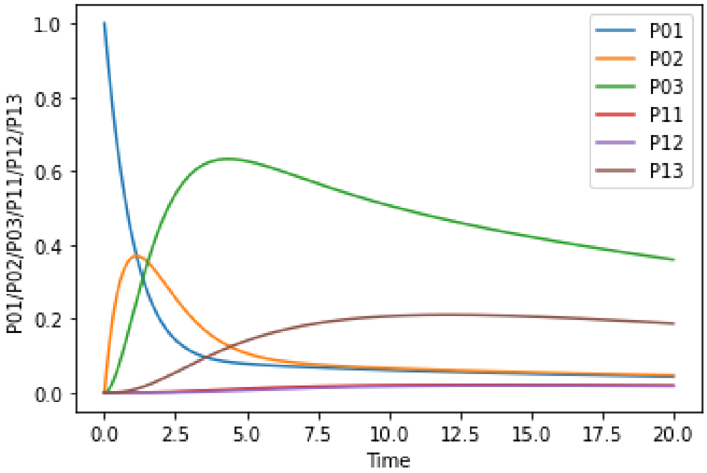


Figure 5.2: Displays three-state probabilities, with corresponding birth and death rates of $\lambda1 = 0.9$, $\lambda2 = 0.9$ $\delta = 0.1$, $\nu = 0.1$, and $\gamma = 0.1$.

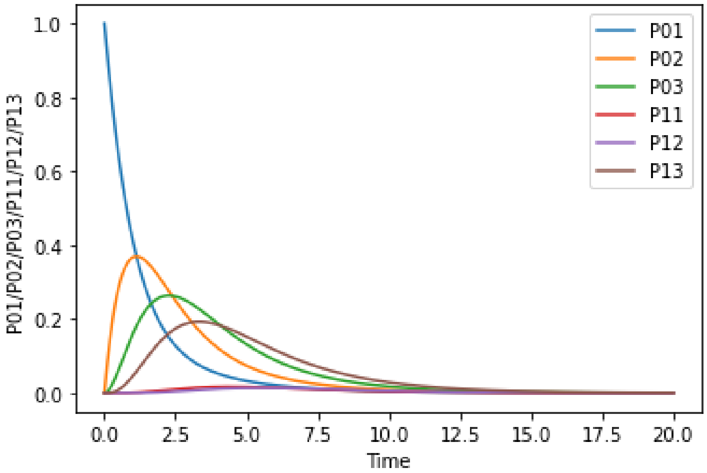


Figure 5.3: Displays three-state probabilities, with corresponding birth and death rates of $\lambda1 = 0.9$, $\lambda2 = 0.9$ $\delta = 0.1$, $\nu = 0.9$, and $\gamma = 0.1$.

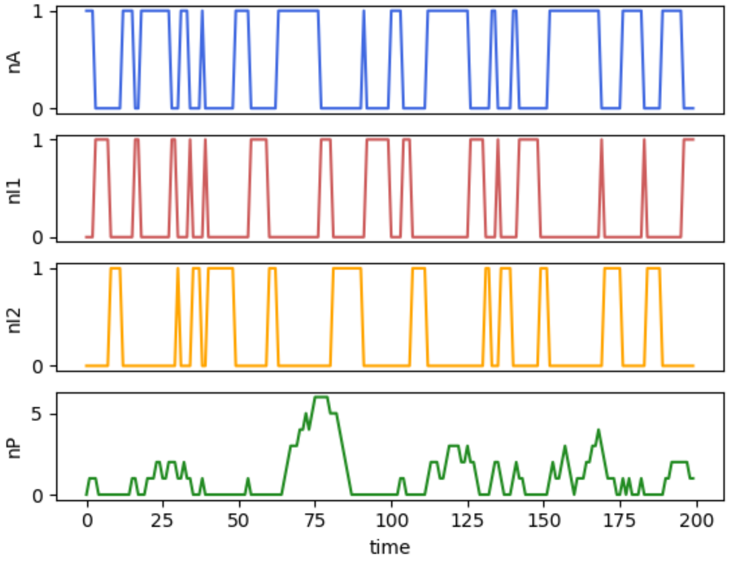


Figure 5.4: Monte Carlo simulation results for one gene that switches back and forth from inactive to active states and produces mRNA. Has corresponding birth and death rates of $\lambda 1 = 0.2$, $\lambda 2 = 0.8$ $\delta = 0.1$, $\nu = 0.7$, and $\gamma = 0.5$.

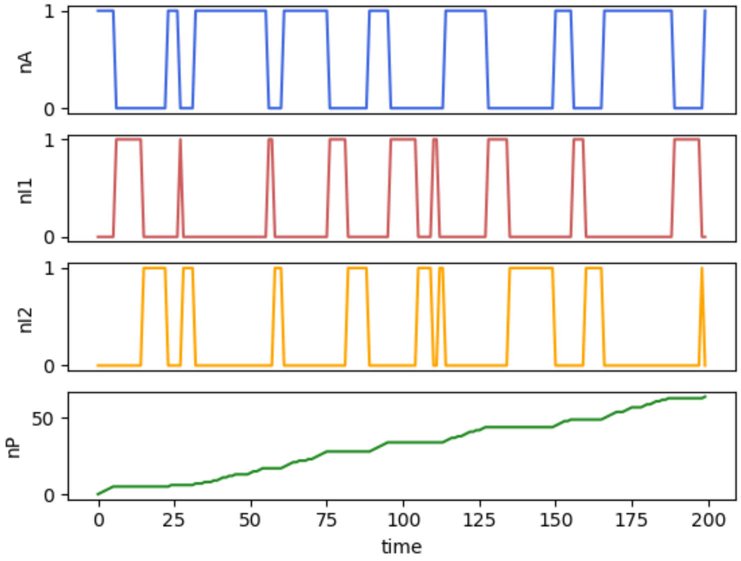


Figure 5.5: Monte Carlo simulation results for one gene that switches back and forth from inactive to active states and produces mRNA. Has corresponding birth and death rates of $\lambda 1 = 0.2$, $\lambda 2 = 0.8$ $\delta = 0.1$, $\nu = 0.7$, and $\gamma = 0.0$.

5.4 Gillespie Algorithm

In this paper we have focused on the increasing complexity of models that mirror microbiological behavior. However, even multi-state models can have difficultly calculating projected behavior as more individual elements or states are added. Models often have either unsolvable or continuous deterministic solutions that fail to capture the discrete behavior by which many processes operate. An example of this being increasing complexity in the ordinary differential equations that provide a probabilistic projection of mRNA production and the bursting activity it can exhibit. It is for these scenarios that the *Gillespie Algorithm* is most applicable[5]. These simulations rely upon sampling the probability distribution given by a governing master equation to a satisfactory point by which macroscopic behavior of the system can be captured. Computationally, this method could be compared to many circuit driven projects that continuously toggle with the question of what sampling frequency will display accurately the behavior of the electrical signals generated. As aforementioned, it is also applicable for biological process which can become analytically unsolvable on a continuous scale or generate discrete data that a continuous model is sought for[5].

Chapter 6

Conclusions

This thesis illustrates three increasing complex stocastic models of mRNA transcription. In theory, the three-state, and mutli-state models would best describe the behavior of many genes as transcription is often highly regulated. These models could be utilized to assess how different rate changes or additional regulation activity could affect mRNA production. While the current arbitrary cases can be further assessed for different rate values, a major improvement to the model would be compiling biologically accurate birth, death, activation, and inactivation rates for genes transcribed by eukaryotic cells. Modifying the models to move from arbitrary time and rate values to biologically accurate units would follow. Next, the models could be examined for their accuracy in predicting mRNA output and if verified or improved upon could then be utilized to predict what would occur when cellular conditions are altered. Additional rate-limiting constrained, such as availability of base pairs for mRNA production or accounting for mRNA processing in a different matter, could also be incorporated to the multi-state model.

Bibliography

[1] Beaudin, L. (n.d.). A Review of the Potts Model.

[2] Chen, J., and Jiao, F. (2022). A Novel Approach for Calculating Exact Forms of mRNA Distribution in Single-Cell Measurements. Mathematics, 10(1), 27

[3] Gibbs J W 1902 *Elementary Principles in Statistical Mechanics* (New York: Scribner)

[4] Guisoni et al. (2018) *Modeling Active Cell Movement With the Potts Model* Frontiers in Physic, vl.6. https://www.frontiersin.org/article/10.3389/fphy.2018.00061

[5] Justin Bois and Micheal Elowitz. (2019). Basic Gillespie Stochastic simulation. Creative Commons Attribution License CC-BY 4.0.

[6] Klindziuk, A. (2021). Stochastic Modeling of DNA Transcription and Gene Expression [Thesis, Rice University]. https://scholarship.rice.edu/handle/1911/111230

[7] Krapivsky P L, Redner A and Ben-Naim E 2010 *A Kinetic View of Statistical Physics* (Cambridge: Cambridge University Press)

[8] Lee, T. I., and Young, R. A. (2013). Transcriptional Regulation and its Misregulation in Disease. Cell, 152(6), 1237–1251. https://doi.org/10.1016/j.cell.2013.02.014

[9] Lodish H, Berk A, Kaiser C A, Krieger M, Bretscher A, Ploegh H, Amon A and Martin K C 2016 *Molecular Cell Biology* 8th edn (W.H. Freeman and Company, N.Y.).

[10] Margulies. E. (2022). Transcription. National Human Genome Research Institute. https://www.genome.gov/genetics-glossary/Transcription

[11] D. A. Mazilu, I. Mazilu, H. T. Williams, “From Complex to Simple: Interdisciplinary Stochastic Models”, IOP Science, Morgan and Claypool Publishers (2018), online ISBN 978-1-64327-120-0, print ISBN 978-1-64327-117-0

[12] Novozhilov, A. S., Karev, G. P., and Koonin, E. V. (2006). Biological applications of the theory of birth-and-death processes. Briefings in Bioinformatics, 7(1), 70–85. https://doi.org/10.1093/bib/bbk006

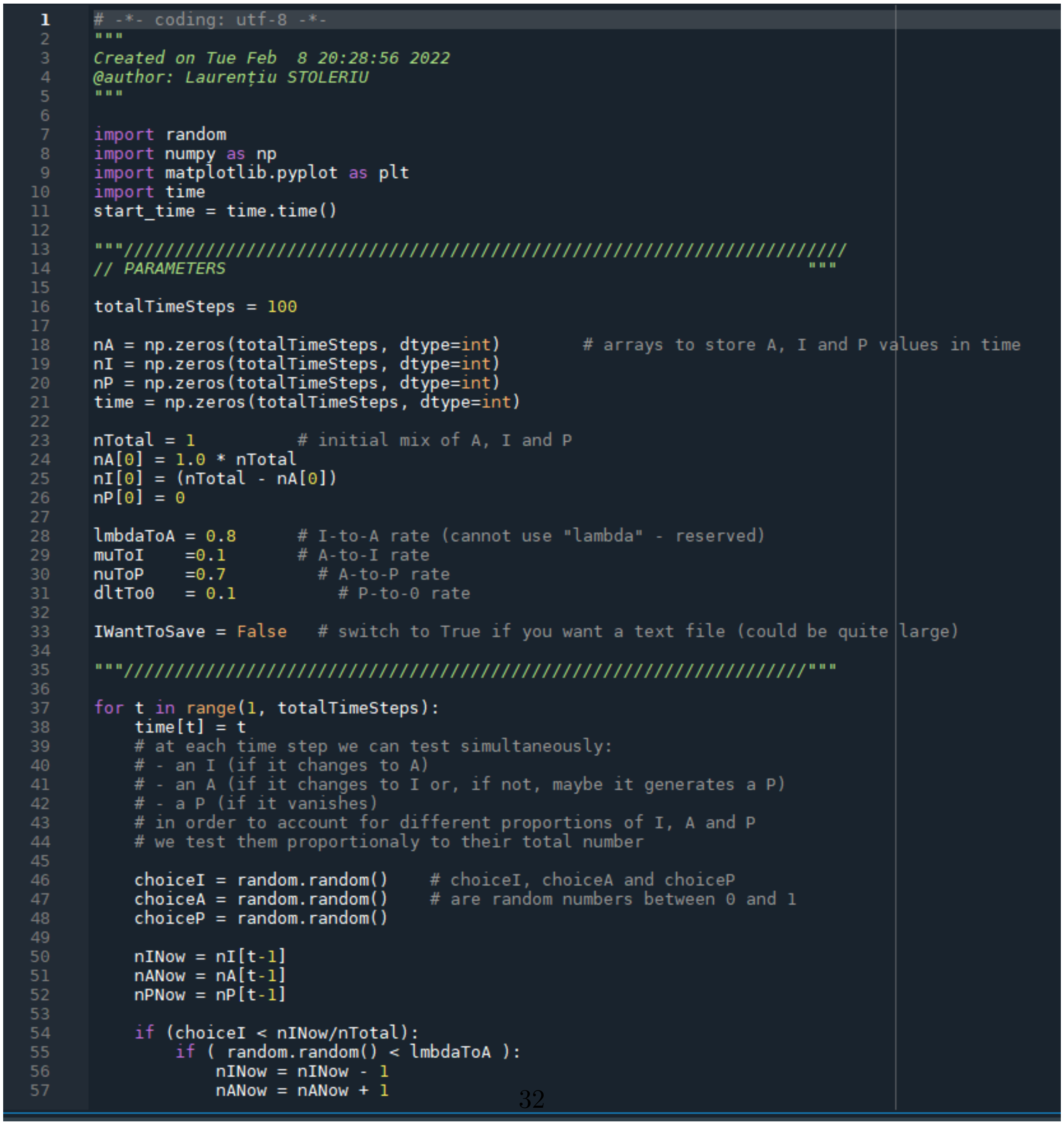
[13] Peccoud, J., and Ycart, B. (1995). Markovian Modeling of Gene-Product Synthesis. Theoretical Population Biology, 48(2), 222–234. https://doi.org/10.1006/tpbi.1995.1027

[14] Van Kampen N G 2002 *Stochastic Processes in Physics and Chemistry* 3rd edn (Elsevier)

[15] Zhou, T., and Zhang, J. (2012). Analytical results for a multistate gene model, SIAM Journal on Applied Mathematics, 72(3), 789–818.

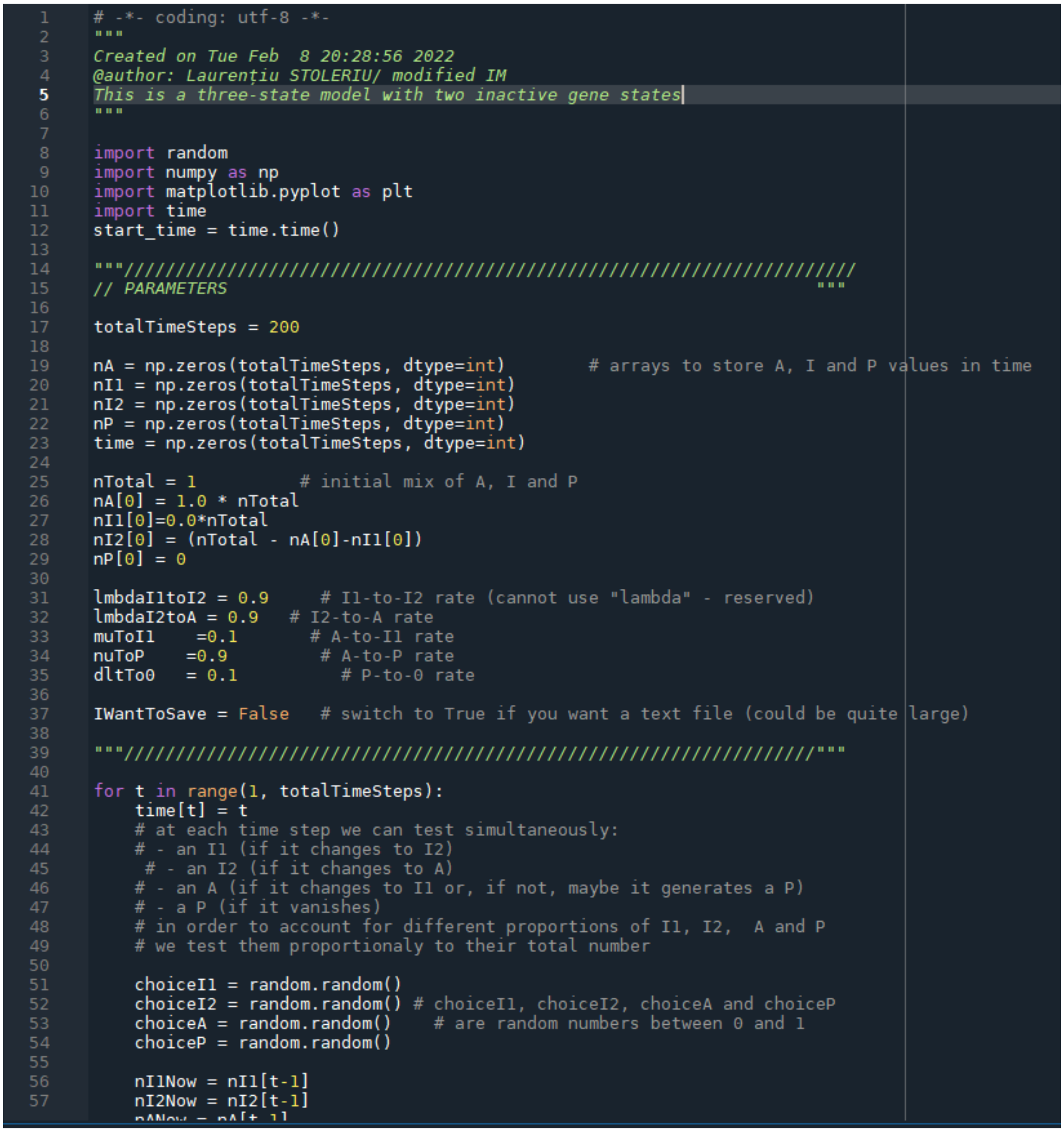
Chapter 7

Appendix



```python
# -*- coding: utf-8 -*-
"""
Created on Tue Feb  8 20:28:56 2022
@author: Laurențiu STOLERIU
"""

import random
import numpy as np
import matplotlib.pyplot as plt
import time
start_time = time.time()

"""////////////////////////////////////////////////////////////////////////
// PARAMETERS                                                          """

totalTimeSteps = 100

nA = np.zeros(totalTimeSteps, dtype=int)        # arrays to store A, I and P values in time
nI = np.zeros(totalTimeSteps, dtype=int)
nP = np.zeros(totalTimeSteps, dtype=int)
time = np.zeros(totalTimeSteps, dtype=int)

nTotal = 1          # initial mix of A, I and P
nA[0] = 1.0 * nTotal
nI[0] = (nTotal - nA[0])
nP[0] = 0

lmbdaToA = 0.8      # I-to-A rate (cannot use "lambda" - reserved)
muToI    =0.1       # A-to-I rate
nuToP    =0.7         # A-to-P rate
dltTo0   = 0.1          # P-to-0 rate

IWantToSave = False   # switch to True if you want a text file (could be quite large)

"""////////////////////////////////////////////////////////////////////"""

for t in range(1, totalTimeSteps):
    time[t] = t
    # at each time step we can test simultaneously:
    # - an I (if it changes to A)
    # - an A (if it changes to I or, if not, maybe it generates a P)
    # - a P (if it vanishes)
    # in order to account for different proportions of I, A and P
    # we test them proportionaly to their total number

    choiceI = random.random()    # choiceI, choiceA and choiceP
    choiceA = random.random()    # are random numbers between 0 and 1
    choiceP = random.random()

    nINow = nI[t-1]
    nANow = nA[t-1]
    nPNow = nP[t-1]

    if (choiceI < nINow/nTotal):
        if ( random.random() < lmbdaToA ):
            nINow = nINow - 1
            nANow = nANow + 1
```

Figure 7.1: Set up of Monte Carlo Simulation code for two-state model.



```python
# -*- coding: utf-8 -*-
"""
Created on Tue Feb  8 20:28:56 2022
@author: Laurențiu STOLERIU/ modified IM
This is a three-state model with two inactive gene states
"""

import random
import numpy as np
import matplotlib.pyplot as plt
import time
start_time = time.time()

"""///////////////////////////////////////////////////////////////////////
// PARAMETERS                                                         """

totalTimeSteps = 200

nA = np.zeros(totalTimeSteps, dtype=int)        # arrays to store A, I and P values in time
nI1 = np.zeros(totalTimeSteps, dtype=int)
nI2 = np.zeros(totalTimeSteps, dtype=int)
nP = np.zeros(totalTimeSteps, dtype=int)
time = np.zeros(totalTimeSteps, dtype=int)

nTotal = 1          # initial mix of A, I and P
nA[0] = 1.0 * nTotal
nI1[0]=0.0*nTotal
nI2[0] = (nTotal - nA[0]-nI1[0])
nP[0] = 0

lmbdaI1toI2 = 0.9     # I1-to-I2 rate (cannot use "lambda" - reserved)
lmbdaI2toA = 0.9   # I2-to-A rate
muToI1    =0.1       # A-to-I1 rate
nuToP    =0.9         # A-to-P rate
dltTo0   = 0.1          # P-to-0 rate

IWantToSave = False   # switch to True if you want a text file (could be quite large)

"""///////////////////////////////////////////////////////////////////"""

for t in range(1, totalTimeSteps):
    time[t] = t
    # at each time step we can test simultaneously:
    # - an I1 (if it changes to I2)
     # - an I2 (if it changes to A)
    # - an A (if it changes to I1 or, if not, maybe it generates a P)
    # - a P (if it vanishes)
    # in order to account for different proportions of I1, I2,  A and P
    # we test them proportionaly to their total number

    choiceI1 = random.random()
    choiceI2 = random.random() # choiceI1, choiceI2, choiceA and choiceP
    choiceA = random.random()    # are random numbers between 0 and 1
    choiceP = random.random()

    nI1Now = nI1[t-1]
    nI2Now = nI2[t-1]
    nANow = nA[t-1]
```

Figure 7.2: Set up of Monte Carlo Simulation code for three-state model.