Introduction to Mathematical Modelling in Synthetic Biology
Version 1

Imperial College London IGEM 2020
https://2020.igem.org/Team:Imperial_College

October 25, 2020

Contents

Chapter 1

Introduction

This brief guide is an introduction to the basic concepts of mathematical modelling in synthetic biology. We will establish a foundation for both deterministic and stochastic approaches, then use these concepts to develop basic models. The limitations and assumptions will also be covered in depth. From here, readers should be able to approach more advanced resources to develop their models further with additional complexity. Awareness of integration, differentiation and high school chemistry is assumed.

Accompanying this document is a tutorial series, which covers the content in video format. Click *here* to view. There is also model code available on our wiki page *here*.

This is version 1 of the package, developed by the 2020 Imperial College IGEM team. It was developed in conjunction with a mentoring program of five other 2020 IGEM teams. Their feedback and insight into the struggles they were facing was invaluable in the development of the package.

By including the input and feedback of a range of potential end-users, we feel this package is thorough and validated. However, we recognise there is certainly room for expansion. We have therefore included the source LaTeX file on our wiki page *here* for future IGEM teams to expand upon as part of their own Human Practises. We hope this package will continue to develop until it becomes a truly comprehensive guide.

Chapter 2

Deterministic Concepts

2.1 Differential Equations

A differential equation (DE) is an equation which contains derivatives of some dependent variable. Differential equations come in many forms, and are an essential concept for much of maths and science. In modelling the dynamics of a system where the variables vary with time, we can form DEs to model this variation. Cell processes such as transcription and translation take time, so we have to describe the variation of mRNA and protein concentrations using differential equations, then attempt to solve those equations.

2.1.1 Characterisation

The properties of differential equations and their solutions can be analysed based on their characterisation. While a deep understanding is not necessary for numerical integration, a basic understanding of these categories is useful. The following terms are used to describe a DE:

1. *Ordinary* differential equations (ODEs) have a dependent variable (and its derivatives) which is only dependent on one independent variable. *Partial* differential equations (PDEs) on the other hand have a dependent variable which is dependent on two or more independent variables. PDEs are a lot more complex to solve.

2. The *order* of a DE is the order of the highest derivative present.

3. *Homogeneous* means every term contains the dependent variable or its derivatives, while *inhomogeneous* implies there is at least one term which does not.

4. *Linear* implies every term is of the form *coefficient* × *derivative*, such that the dependent variable and its derivatives are never in any function.

5. *Coupled* implies there are two or more DEs in which they share two or more dependent variables. In this way, the variation of the dependent variable of one DE is dependent on the variation of the dependent variable of another DE.

2.1.2 Solutions

The solutions of a DE are functions which satisfy the equation such that no derivatives are present in the solution. A lot of the time, DEs can't be solved analytically (that is, a general algebraic solution). This is because integration is sometimes impossible for more complex equations (such as the Hill equation, in general). However, they can be solved numerically using iterative computational methods. There are many different methods for different situations, but programs like MatLab or packages like Scipy in Python have functions which do the hard work.

2.2 Mass Action Kinetics (MAK)

Mass action kinetics is a framework to analyse chemical reactions based on the assumption that the rate of reactions can be modelled by the concentration of the reactants. There is also a coefficient of proportionality to be found empirically. Once these coefficients are known, the rate equation (a DE) can be solved to provide a model to describe how concentrations vary with time.

2.2.1 A Simple Reaction

Consider the reaction

$$A \to C. \tag{2.1}$$

For every one A that reacts, one C is outputted as the product. That is, for every reaction, one C is created and one A is used up. Given C doesn't react, the rates won't be dependent on C.

In this equation, we would model the rates as follows:

$$\frac{d[A]}{dt} = -k[A] \tag{2.2}$$

and

$$\frac{d[C]}{dt} = k[A]. \tag{2.3}$$

As A is used up, the concentration decreases. As C is produced, the concentration increases. Since the rate of production of C is equivalent to the rate of A being used up, the equations are very similar.

2.2.2 Reverse Reaction

Suppose we introduce the reverse reaction, so C can react and go back to A:

$$A \rightleftharpoons C. \tag{2.4}$$

This would mean we effectively have two reactions, A to C and C to A. The new rates are the sum of these two ODEs. Therefore, in total, we have

$$\frac{d[A]}{dt} = -k_1[A] + k_2[C] \tag{2.5}$$

and

$$\frac{d[C]}{dt} = k_1[A] - k_2[C]. \tag{2.6}$$

2.2.3 Two Reactants

Let's further the complexity, and add an additional reactant:

$$A + B \rightleftharpoons C. \tag{2.7}$$

Intuitively, we see that an A and a B must react to make a C. This means that if there are no A, no reaction will occur. This implies rate is dependent on concentration of A and B in one combined term. Likewise, the reverse reaction is only dependent on C, and produces an A and a B. We therefore have 3 rates here:

$$\frac{d[A]}{dt} = -k_1[A][B] + k_2[C], \tag{2.8}$$

$$\frac{d[B]}{dt} = -k_1[A][B] + k_2[C] \tag{2.9}$$

and

$$\frac{d[C]}{dt} = k_1[A][B] - k_2[C]. \tag{2.10}$$

2.2.4 Reaction Coefficients

Consider the reaction

$$2A \rightleftharpoons C. \tag{2.11}$$

Referencing section C, we observe that the rates will be given as

$$\frac{d[A]}{dt} = -k_1[A][A] + k_2[C] = -k_1[A]^2 + k_2[C] \tag{2.12}$$

and

$$\frac{d[C]}{dt} = k_1[A][A] - k_2[C] = k_1[A]^2 - k_2[C]. \tag{2.13}$$

So the *coefficient* in the reaction equation will be the *power* in the rate equation.

2.2.5 Degradation

In some instances, a molecule may degrade to a some other species which isn't relevant to the system in question, and cannot revert back to the original molecule. We write this as

$$A \rightarrow \varnothing. \tag{2.14}$$

This is modelled just as in part A, but there is no product. Instead, we ignore it, since we only care about certain species within a system. The rate coefficient for degradation is typically given as δ, giving the rate equation

$$\frac{d[A]}{dt} = -\delta[A]. \tag{2.15}$$

A good understanding of Mass Action Kinetics will be very useful for appreciating exactly what the mathematical model of chemical reactions is describing.

2.2.6 Exercise 2.2

Given the chemical equation for the production of hydrogen chloride from hydrogen and chloride find the correct rate equation for hydrogen chloride (given that the forward reaction rate is equal to K_{F} and the backward reaction rate is equal to K_{B}):

$$H_2 + Cl_2 \rightleftharpoons 2HCl. \tag{2.16}$$

$$A.)\frac{d[HCl]}{dt} = +K_{\mathrm{F}}[H_2][Cl_2] - K_{\mathrm{B}}[HCl]^2 \tag{2.17}$$

$$B.)\frac{d[HCl]}{dt} = +K_{\mathrm{F}}[H_2][Cl_2] - K_{\mathrm{B}}[HCl] \tag{2.18}$$

$$C.)\frac{d[HCl]}{dt} = +K_{\mathrm{F}}[H]^2[Cl]^2 - K_{\mathrm{B}}[HCl]^2 \tag{2.19}$$

2.3 Hill Equation

In any reaction where two species bind together, they may unbind and return to their constituent parts. This is the same form as in section 2.2.3, but where the larger species may have multiple binding sites, such that we have a reaction equation of the form

$$A + n * B \rightleftharpoons C, \tag{2.20}$$

where A is the larger species, B is the smaller species, C is the complex formed from the binding and n is the number of binding sites on A. For example, binding of a ligand to a macromolecule, a transcription factor to a promoter, or a substrate to an enzyme. Using MAK, we can derive an equation called the Hill Equation from the rate equation, based on a few assumptions. While it has limitations in its accuracy, the Hill Equation is a good starting point when the dynamics of the binding is unknown.

2.3.1 Derivation

We define the forward reaction constant k_a for association and the reverse reaction constant k_d for dissociation. Considering the ODE for the complex using MAK, we therefore have

$$\frac{d[C]}{dt} = k_\mathrm{a}[A][B]^n - k_\mathrm{d}[C]. \tag{2.21}$$

If [C] is initially zero, the concentrations will shift until the forward reaction rate is equal to the reverse reaction rate, such that the concentrations no longer vary with time. We call this point equilibrium. This means at equilibrium, the concentration of the complex won't vary with time. We therefore equate this equation to zero:

$$\frac{d[C]}{dt} = k_\mathrm{a}[A][B]^n - k_\mathrm{d}[C] = 0. \tag{2.22}$$

Rearranging, we define the apparent dissociation constant, $\mathrm{K}_{\mathrm{d},}$ as the ratio of the dissociation rate constant and the association rate constant.

$$K_\mathrm{d} = \frac{k_\mathrm{d}}{k_\mathrm{a}} = \frac{[A][B]^n}{[C]}. \tag{2.23}$$

We can then calculate the proportion of bound A molecules to the proportion of total A molecules (bound and unbound), which we call θ:

$$\theta = \frac{\text{Bound A Molecules}}{\text{Total A molecules}} = \frac{[C]}{[A]+[C]}, \tag{2.24}$$

remembering that A is the larger species.

Using equation (2.23), we can substitute $[C] = \frac{[A][B]^n}{K_\mathrm{d}}$ and rearrange to get the equation

$$\theta = \frac{\frac{[A][B]^n}{K_\mathrm{d}}}{[A] + \frac{[A][B]^n}{K_\mathrm{d}}} = \frac{[A][B]^n}{K_\mathrm{d}[A] + [A][B]^n} = \frac{[B]^n}{K_\mathrm{d} + [B]^n}. \tag{2.25}$$

This is the proportion of the Complex molecule (bound A) to the total number of A molecules in terms of their concentrations. It is also possible to find the proportion of unbound A molecules, simply by calculating (1-θ)

2.3.2 Limitations

- Assumption:

  Assumes all Bs bind to A simultaneously, as opposed to in succession over some time.

  Does not account for the interaction between successive Bs and their cooperativity (this models assumes very positive cooperativity).

  Does not consider partial binding (ie not all binding sites filled), and the functionality of this partially bound complex.

  Justification:

  This model requires very little empirical knowledge to be implemented, so is favourable as a starting point when knowledge of the dynamics and properties of the species is limited.

- Assumption:

  The reaction will take time to reach equilibrium. When using this equation in DEs, we assume this binding equilibrium is reached instantaneously.

  Justification:

  In the context of cell processes, the timescale of reaching equilibrium is much smaller for binding. We therefore make what's called a *quasi-stationary approximation* that since this binding reaction is much faster, steady state is reached instantaneously.

- Assumption

  The system is well mixed, or alternatively diffusion is instantaneous.

  Justification

  For liquid solutions, this assumption is valid. However, cells aren't quite liquid solutions - they are more viscous, meaning diffusion is difficult and slow. There are also different compartments within a cell. However, modelling the compartmentalisation of the cell is much more advanced (although necessary for a complete whole-cell model).

- Assumption

  The large molecule concentration, A, is much smaller than the small molecule concentration, B. In this model, we don't consider whether the concentration of smaller molecules B is a limiting factor.

  Justification

  This assumption is necessary for the model, so ensure this is the case before using the hill equation. If this isn't the case, the binding process should be modelled using a DE.

The hill equation is a very common equation used in biochemistry. While not encompassing all the dynamics of a binding process, it is a good starting point to build up from.

Alternatively, when creating the simulation, it is possible to define the binding and unbinding process as a differential equation, meaning quasi-stationary assumption isn't an assumption that is made. However, more information about the dynamics is required (k_a and k_d must both be known).

To explore the bounds of these assumptions and when they are valid and invalid, see coding challenge 3.1.1.

2.3.3 Example 1: Macromolecule and Ligand Binding

Consider a simple binding of a large macromolecule (eg FUTt, a protein) with ligands (eg iron ions). The reaction equation is

$$Macromolecule + n * Ligand \rightleftharpoons Complex, \tag{2.26}$$

where n is the number of ligand binding sites in the macromolecule.

Suppose we want to find the concentration of the complex formed from their binding at equilibrium. The Hill equation describes the proportion of bound macromolecules. Therefore multiplying this by the *initial* macromolecule concentration will give the concentration of the complex formed:

$$[C] = [M_{\text{initial}}]\theta = [M_{\text{initial}}](\frac{[L]^n}{K_{\text{d}} + [L]^n}) \tag{2.27}$$

where [C] is the concentration of the complex, $[M_{\text{initial}}]$ is specifically the initial concentration of unbound macromolecule *before* any binding (or alternatively just the total concentration of M both bound and unbound, $[M]_0$), $[L]$ is the concentration of the ligand, K_{d} is the dissociation constant and n is the hill coefficient.

Alternatively, it is possible to find the concentration of unbound macromolecules *after* binding. θ gives the proportion of bound macromolecules in the complex, so naturally (1-θ) gives the proportion of unbound macromolecules. Therefore we have the alternative equation

$$[M_{\text{unbound}}] = [M_{\text{initial}}](1 - \frac{[L]^n}{K_{\text{d}} + [L]^n}) \tag{2.28}$$

2.3.4 Example 2: Transcription Rate

With a weak promoter, the binding of the transcription factor to the promoter is not perfect, meaning there is some dissociation. This is given as

$$Promoter + n * TF \rightleftharpoons BoundPromoter. \tag{2.29}$$

Therefore the hill equation for this binding is given as

$$\theta = \frac{[TF]^n}{K_{\text{d}} + [TF]^n}. \tag{2.30}$$

If the transcription factor *activates* the transcription of the downstream gene, then only the bound promoters will transcribe the downstream gene. Therefore if we multiply θ by the maximum transcription rate k_{TX} if all promoters were activated, we get the transcription rate:

$$\frac{d[mRNA]}{dt} = k_{\text{TX}}\theta = k_{\text{TX}}\frac{[TF]^n}{K_{\text{d}} + [TF]^n}. \tag{2.31}$$

On the other hand, if the transcription factor *represses* transcription of the downstream gene, we multiply k_{TX} by the proportion of unbound promoters, given by (1-θ):

$$\frac{d[mRNA]}{dt} = k_{\text{TX}}(1-\theta) = k_{\text{TX}}(1 - \frac{[TF]^n}{K_{\text{d}} + [TF]^n}). \tag{2.32}$$

2.3.5 Example 3: Michaelis–Menten Equation

Enzyme reactions are initiated by the binding of the substrate and the enzyme, which is a reversible process. This in turn releases a product, and regenerates the original enzyme. Therefore we have the reaction equation

$$E + S \rightleftharpoons ES \rightarrow E + P, \tag{2.33}$$

where E is the enzyme, S is the substrate, ES is the complex and P is the product. We define the rate constants as k_{a} and k_{d} for the association and dissociation rate constants as we have previously and k_{cat} for the catalytic rate constant. Just as before, we can define the proportion of complexes using the hill equation:

$$\theta = \frac{[S]}{K_{\text{d}} + [S]}. \tag{2.34}$$

If we then multiply this by the total concentration of enzymes (both bound and unbound, also given by the initial enzyme concentration), $[\mathrm{E}]_0$, we can find the concentration of the complex:

$$[ES] = [E]_0 \frac{[S]}{K_{\mathrm{d}} + [S]}. \tag{2.35}$$

Given the rate equation for the catalysis is given by

$$\frac{d[P]}{dt} = k_{\mathrm{cat}}[ES], \tag{2.36}$$

we can substitute [ES] for equation 2.35 to give

$$\frac{d[P]}{dt} = k_{\mathrm{cat}}[E]_0 \frac{[S]}{K_{\mathrm{d}} + [S]}. \tag{2.37}$$

This equation is known as the Michaelis-Menton equation. Usually, instead of writing $\mathrm{k}_{\mathrm{cat}}[\mathrm{E}]_0$, we instead write this as $\mathrm{V}_{\mathrm{max}}$, which would be the maximum catalysis rate if all of the enzymes were bound.

2.3.6 Dissociation Constant

In this guide, we derived the apparent dissociation constant at equilibrium, K_{d}, using Mass Action Kinetics. However, the form of this constant in literature often varies. An alternative definition is for the microscopic disassociation constant, K_{A}, sometimes written as K_{D}, such that

$$K_{\mathrm{A}}{}^{n} = K_{\mathrm{D}}{}^{n} = K_{\mathrm{d}} = \frac{k_{\mathrm{d}}}{k_{\mathrm{a}}}. \tag{2.38}$$

This can get quite confusing, given the similar notation. For that reason, care should taken in clarifying exactly which constant is being used for any given model.

K_{D} can be calculated experimentally by calculating the concentrations of each of the species at equilibrium, and using Equation (2.23) (see *here* for more information).

2.3.7 Hill Coefficient

In the derivation of the Hill equation, we found that the Hill coefficient, n, was given as the number of binding sites on the larger molecule. However, we also considered the limitations of this derivation, primarily in that we don't consider successive binding. We therefore don't consider how the large molecule molecule may function as "bound" even with partial binding, and the interaction between subsequent binding and their cooperatively.

However, there is a way to work around these issues, and that is via the hill coefficient. When finding n empirically, it is very rare that the hill coefficient is equivalent to the number of binding sites. For example, the hill coefficient of hemoglobin and oxygen binding was calculated from the binding curve. Despite there being 4 ligand binding sites, n was found to be 1.7–3.2. In this way, we account for both the functionality with partial binding *and* the cooperativity of subsequent binding. With this in mind, the hill coefficient can be considered the interaction coefficient, as oppose to the number of binding sites.

2.3.8 Exercise 2.3a

Calculate θ, the proportion of bound A to total A molecules, in the binding reaction: $A + B \rightleftharpoons C$.

$$A.)\theta = \frac{[C]}{[A] + [C]} \tag{2.39}$$

$$B.)\theta = \frac{[B]}{[A] + [C]} \tag{2.40}$$

$$C.)\theta = \frac{[A]}{[A] + [C]} \tag{2.41}$$

Chapter 3

Deterministic Model Examples

3.1 Binding

We elaborated extensively on the Hill equation as an example of the application of mass action kinetics. We made a quasi-stationary assumption when using it, as well as assuming one of the reagents was far in excess. Here, we take a deep dive into these assumptions by considering both the value given from the Hill equation, as well as the steady state value given by the ODEs. We can visualise this using programming by simulating them both together. We invite readers to attempt the coding challenge below. The modal code will be available for reference.

3.1.1 Coding Challenge

Consider the equation

$$A + B \rightleftharpoons C, \tag{3.1}$$

where the association rate constant $k_{\mathrm{a}} = 0.05 s^{-1} M^{-1}$ and the dissociation rate constant $k_{\mathrm{d}} = 0.00001 s^{-1}$. Initial concentrations for A, B and C are 1, 5 and 0 M respectively. We assume the hill coefficient $n = 1$

Exercises

1. Hill Equation
   (a) Calculate the value of the apparent dissociation rate constant, K_{d}.
   (b) Calculate the hill equation, θ.
   (c) Calculate the expected concentrations for the complex C and the unbound A molecules.
2. ODEs (This will be a lot trickier, since knowledge of ODE solvers is necessary)
   (a) Using mass action kinetics, define the three ODEs necessary to describe the full dynamics of this system.
   (b) In your preferred language, simulate the ODEs by solving them over a time period of 10 seconds.
3. Evaluation
   (a) Plot both the solved ODEs and the value given from the hill equation on the same graph.
   (b) Compare the value the Hill Equation gives to the value the ODEs give for bound and unbound concentrations, A and C, once steady state is reached. Is the approximation accurate?

(c) Considering that cell processes take many minutes or hours, is the quasi-stationary assumption a valid assumption?

(d) Try decreasing the value for k_a. What is the limit at which the reaching of steady state extends beyond 1 minute?

(e) Try varying the values for initial concentrations of A, B and C. Given we assumed B was far in excess, try giving them comparable values, or even making B the limiting factor.

(f) Try varying the hill coefficient n to see how this affects the simulation.

3.2 Basic Gene Expression

This is arguably the most fundamental model for Synthetic biology. Regardless of the complexity of a project, manipulating DNA will affect gene expression. When creating novel parts and systems, we have so many options. For example, we can vary promoter and ribosome binding site (RBS) strength, we can introduce self-cleaving Ribozymes, and we can have genes interacting with one another, such as in the Repressillator system. Here, we introduce the framework in the most simple case: a single composite part with a strong promoter, a strong RBS, a coding sequence (CDS) for the gene we want translated and a terminator.

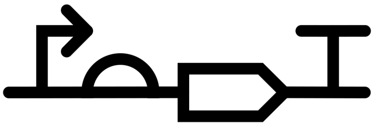

Figure 3.1: Simple Construct

3.2.1 Transcription

Since we have a strong promoter, we don't need to consider the proportion of bound promoters - we assume all are activated (We also assume concentration of transcription factors is much greater that concentration of the engineered gene). The RNA polymerise binds to the promoter and transcribes the downstream gene (the RBS and CDS) at a rate given by the length of the gene in terms of the number of base pairs (We assume RNA polymerase concentration is much greater than gene concentration). For a particular cell, we can then multiply this by the copy number of the gene and divide by the volume of the cell to find the maximum transcription rate, $\mathrm{k_{TX}}$.

Therefore, the rate of transcription is given as

$$\frac{d[mRNA]}{dt} = k_\mathrm{TX} \tag{3.2}$$

3.2.2 mRNA Degradation

mRNA degrades in the cell so as to regulate the amount of protein the mRNA can translate for the sake of resource management. Different mRNAs have variable stability in the cell, with order of magnitude differences in half-life. For example, the addition of a polyA tail increases its stability, with longer polyA tails increasing stability further still. Once degraded, the mRNA has no use, so we form the reaction equation

$$mRNA \rightarrow \varnothing. \tag{3.3}$$

We therefore have the rate equation

$$\frac{d[mRNA]}{dt} = -\delta_{\mathrm{mRNA}}[mRNA]. \tag{3.4}$$

3.2.3 Translation

Once the gene has been transcriped, Ribosomes then bind to the RBS of the mRNA and translate the mRNA, producing the protein which it codes for. We assume a uniform and consistent density of tRNA amino-acid complexes and ribosomes, such that translation rate of a given mRNA is only dependent on the length of the mRNA in terms of the number of bases (encapsulated ib k_{TL}) and the concentration of mRNA ($[mRNA]$). We can therefore simplify the dynamics of translation to the reaction equation

$$mRNA \to mRNA + Protein, \tag{3.5}$$

noting that the mRNA is reproduced after translation. We define the translation rate constant, k_{TL}, which encompasses the complexity of the process. We therefore have the rate equation

$$\frac{d[Protein]}{dt} = k_{\mathrm{TL}}[mRNA]. \tag{3.6}$$

Translation is a discrete event that takes an amount of time. We assume instead that it is a continuous process for the sake of creating our DEs, although we miss some dynamics in doing this, such as an initial delay (nothing will be translated for at least the amount of time needed to transcribe one mRNA), and the dynamics of this binding process (for example, mRNA degradation may vary when being translated).

3.2.4 Protein Dilution

Proteolysis is an extremely slow process if no catalyst is present, taking hundreds of years. In general, we may therefore assume protein degradation is null. However, when cells grow, they share their proteins approximately equally between the daughter cells. This process is an event, as oppose to a continuous process, but for the purpose of developing our DEs, we will assume this process is continuous. For a given cell, protein concentration will decrease according to the reaction equation

$$Protein \to \varnothing \tag{3.7}$$

We define δ_{dilut} as the rate coefficient, dependent of the growth rate of the cells (Note that the engineered gene puts strain and burden on the cell, reducing growth rate). We therefore have the rate equation

$$\frac{d[Protein]}{dt} = -\delta_{\mathrm{Protein}}[Protein]. \tag{3.8}$$

.

3.2.5 Total ODEs

We have now modelled every reaction necessary to describe the rate equations for mRNA and Protein. Putting them all together, we have the total ODES:

$$\frac{d[mRNA]}{dt} = k_{\mathrm{TX}} - \delta_{\mathrm{mRNA}}[mRNA]. \tag{3.9}$$

and

$$\frac{d[Protein]}{dt} = k_{\text{TL}}[mRNA] - \delta_{\text{Protein}}[Protein]. \tag{3.10}$$

The constants can either be empirically measured, found in literature, or estimated based on certain factors. The ODEs can then be solved numerically.

3.2.6 Coding Challenge

Consider the Total ODEs as above, with the following values for the rate coefficients: $k_{\text{TX}} = 0.001M/s$; $\delta_{\text{mRNA}} = 0.001s^{-1}$; $k_{\text{TL}} = 0.002s^{-1}$; and $\delta_{\text{Protein}} = 0.001s^{-1}$. Set initial values to be $0M$ for both mRNA and Protein concentration.

1. First, consider the dynamics of the system. No protein can be produced until there is some mRNA, but once transcribed, protein translates at a faster rate. How do you expect the graphs of concentration against time to appear?

2. Simulate the ODEs over a time period of 2.5 hours and plot them together on a graph of concentration against time.

3. How do the graphs compare to your expectations? Try justifying the key areas of the graph.

4. Try messing around with the rate coefficients. What happens if we set protein dilution to zero? Try simulating 5 different rates, and plot them all on the same graph to visualise the effect it has.

5. Instead of assuming the promoter is strong, try incorporating the hill equation into the transcription rate ODE by referencing 2.3.4. Set [TF] to $0.01M$, $n = 1$ and $k_{\text{d}} = 10^{-6}M$

6. The transcription factor concentration can often be varied. Try plotting protein concentration over time for [TF] of 0, 0.2, 0.4, 0.6, 0.8 and 1.0.

3.3 Repressilator

The repressilator is a system of 3 genes, where the protein produced by one inhibits the transcription of the next gene in a cycle. In the original paper which formualted this system (Elowitz and Leibler, Nature, 2000), an additional gene producing GFP was used to observe this oscillatory behaviour:

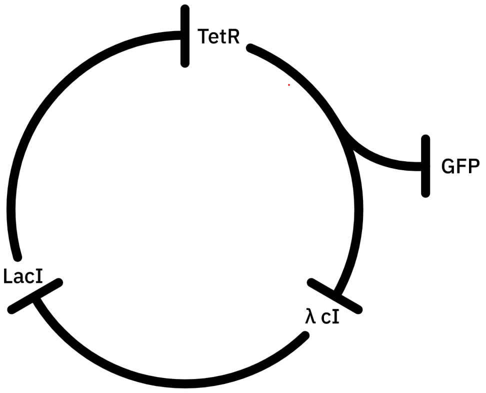

Figure 3.2: Repressilator Diagram with GFP as output

To model this system, we must add complexity to the gene expression model by considering two additional factors: the strength and the leakiness of the promoter.

3.3.1 Adding Complexity

Promoters are activated or repressed in the presence of transcription factors. We initially assumed the promoter was very strong (as in, the binding was very strong), so all promoters were bound. However, this is not generally the case, and we can account for the dissociation of the promoter-TF binding using the hill equation, as described in section 2.3.4. For the repressilator system, the promoters are repressed by our proteins, so we have the generic equation

$$\frac{d[mRNA]}{dt} = k_{\mathrm{TX}}(1 - \frac{[P]^n}{K_{\mathrm{d}} + [P]^n}) = \frac{k_{\mathrm{TX}}}{1 + \frac{[P]^n}{K_{\mathrm{d}}}}. \tag{3.11}$$

Additionally, even when fully bound, the promoter is still able to transcribe the downstream gene, but at a reduced rate. We describe this as the leakiness of the promoter. We therefore redefine the maximum transcription rate, k_{TX}, as $\alpha + \alpha_0$ in the presence of no TF and α_0 when saturated with TF. We therefore redefine the transcription rate as

$$\frac{d[mRNA]}{dt} = \frac{\alpha}{1 + \frac{[P]^n}{K_{\mathrm{d}}}} + \alpha_0. \tag{3.12}$$

3.3.2 Total ODEs

Using the basic gene expression model for the rest of the dynamics, we have the total equations for mRNA and Protein. Putting them all together, we have the total ODES:

$$\frac{d[mRNA_{\mathrm{i}}]}{dt} = \frac{\alpha_{\mathrm{i}}}{1 + \frac{[P_{\mathrm{j}}]^n}{K_{\mathrm{d_i}}}} + \alpha_{0_{\mathrm{i}}} - \delta_{\mathrm{mRNA_i}}[mRNA_{\mathrm{i}}]. \tag{3.13}$$

and

$$\frac{d[P_{\mathrm{j}}]}{dt} = k_{\mathrm{TL_j}}[mRNA_{\mathrm{i}}] - \delta_{\mathrm{Protein_j}}[P_{\mathrm{j}}]. \tag{3.14}$$

where $i = (Protein1, Protein2, Protein3)$ and $j = (Protein3, Protein1, Protein2)$. In the Elowitz paper, these were LacI, TetR and λ cI.

3.3.3 Non-Dimensionalisation

In a process known as non-dimensionalisation, we can change the units, so as to be closer to the values we can empirically measure. In the original paper, the units were quite different from SI (from the french *Système international d'unités*, this gives the standard units for all dimensions, from time to distance to electrical charge to concentration): time is rescaled in units of mRNA lifetime; protein concentration in units of K_{d}; mRNA concentrations in units of translation efficiency (average number of proteins produced per mRNA molecule); β is defined as the ratio of protein decay rate to mRNA decay rate. In this way, the ODEs can be written as

$$\frac{d[mRNA_{\mathrm{i}}]}{dt} = \frac{\alpha_{\mathrm{i}}}{1 + [P_{\mathrm{j}}]^n} + \alpha_{0_{\mathrm{i}}} - [mRNA_{\mathrm{i}}]. \tag{3.15}$$

and

$$\frac{d[P_{\mathrm{j}}]}{dt} = -\beta([P_{\mathrm{j}}] - [mRNA_{\mathrm{i}}]). \tag{3.16}$$

A deep understanding of this non-dimensionalisation is not completely necessary, although an awareness is essential. When researching literature or searching databases for constants, be careful with the units, as they may not be SI.

3.3.4 Coding Challenge

For this coding challenge, we use the equations we derived using SI units, as oppose to the equations in the original paper. We first set the constants as follows:

$\alpha_1 = \alpha_2 = \alpha_3 = 0.001M/s$;
$\alpha_{0_1} = \alpha_{0_2} = \alpha_{0_3} = 10^{-6}M/s$;
$K_{\mathrm{d}_1} = K_{\mathrm{d}_1} = K_{\mathrm{d}_1} = 10^{-6}M^n$;
$n_1 = n_2 = n_3 = 2$;
$\delta_{\mathrm{mRNA}_1} = \delta_{\mathrm{mRNA}_2} = \delta_{\mathrm{mRNA}_3} = 0.001s^{-1}$;
$k_{\mathrm{TL}_1} = k_{\mathrm{TL}_2} = k_{\mathrm{TL}_3} = 0.002s^{-1}$;
$\delta_{\mathrm{Protein}_1} = \delta_{\mathrm{Protein}_2} = \delta_{\mathrm{Protein}_3} = 0.001s^{-1}$.

Set initial values to be $0.1M$ for *mRNA1* and *Protein1*, and $0M$ for *mRNA2*, *Protein2* , *mRNA3* and *Protein3*.

1. Define the above constants in your code.
2. Define the 6 ODEs to solve.
3. Solve and plot the 3 protein concentrations over a time period of 10 hours.
4. Experiment with changing the values of some of the variables. What happens if the hill coefficient is decreased? What about if one of the promoters is particularly leaky (large a_0)?

3.4 Auto-Activation and Auto-Inhibition

The auto-activation and auto-inhibition systems are gene circuits where the transcription factor which regulates the promoter is the same as the protein it encodes for. In this way, the gene circuit is self-regulating.

3.4.1 Auto-Activation

Here, the protein being produced by the gene also activates transcription, meaning even a small amount of a protein will initiate a positive feedback loop, producing a lot of the protein.

The model is therefore written as

$$\frac{d[mRNA]}{dt} = k_{\mathrm{TX}} \frac{[Protein]^n}{K_{\mathrm{d}} + [Protein]^n} - \delta_{\mathrm{mRNA}}[mRNA]. \tag{3.17}$$

and

$$\frac{d[Protein]}{dt} = k_{\mathrm{TL}}[mRNA] - \delta_{\mathrm{Protein}}[Protein]. \tag{3.18}$$

In an ideal world, this would make for an excellent diagnostics or detection tool. Any amount of a protein, even if it is a single molecule, would eventually reach steady state with a large concentration of the protein. However, in reality, promoters aren't 100% dependent on transcription factor concentration - they are leaky. Therefore, such a system would be susceptible to false positives. This system still has its uses however.

3.4.2 Auto-Inhibition

Here, the protein produced inhibits transcription, meaning in high concentrations of protein, no more of the protein is produced. In cells, concentration would then decrease through dilution. In this way, transcription is restricted, so there is a limit to how much protein can be produced through translation.

The model is written as

$$\frac{d[mRNA]}{dt} = k_{\text{TX}}(1 - \frac{[Protein]^n}{K_{\text{d}} + [Protein]^n}) - \delta_{\text{mRNA}}[mRNA]. \tag{3.19}$$

and

$$\frac{d[Protein]}{dt} = k_{\text{TL}}[mRNA] - \delta_{\text{Protein}}[Protein]. \tag{3.20}$$

3.4.3 Coding Challenge

Consider the above models with the following constants: $k_{\text{TX}} = 0.001M/s$; $k_{\text{d}} = 10^{-8}M^n$; $n = 1$; $\delta_{\text{mRNA}} = 0.001s^{-1}$; $k_{\text{TL}} = 0.002s^{-1}$; and $\delta_{\text{Protein}} = 0.001s^{-1}$.

Auto-Activation

Set initial mRNA and protein concentration to be $0M$ and $10^{-7}M$ respectively.

1. Define the constants and ODEs in your code
2. Solve and plot mRNA and protein concentration over a time period of 1 minute
3. Solve and plot over a time period of 2 hours
4. Compare the two graphs - does this behaviour make sense?
5. Vary initial protein concentration to observe how steady state is always eventually reached, no matter how small protein concentration is (noting that this is because our model is continuous, not discrete).
6. Incorporate leakiness into the model (add a_0 with a value of $10^{-5}M/s$). How does this affect the accuracy of the model? For example, plot graphs of varying initial protein concentration. At what magnitude do the lines become indistinct from one another?

Auto-Inhibition

Set initial mRNA and protein concentration to be $0M$ and $0.0001M$ respectively.

1. Define the constants and ODEs in your code
2. Solve and plot mRNA and protien concentration over a time period of 2 hours
3. Vary initial protein concentration, both increasing and decreasing the order of magnitude. Is there a trend in the steady state reached?
4. Incorporate leakiness into the model (add a_0 with a value of $10^{-5}M/s$). How does this affect the behaviour?

Chapter 4

Stochastic Concepts

So far, we have learnt about ODEs and applied them in deterministic models of gene expression and simple genetic circuits. An implicit assumption of deterministic models is that all variables and parameters vary in a continuous, non-random fashion. This section deals with the scenario where this assumption is not upheld. In doing so, we must consider theory of statistical analysis, which can get quite complex.

4.1 Random event

Random events are events influenced by chance. Chance and its properties are studied by random experiments, for example: to roll a die, to flip a coin. The outcome is each possible result of the experiment and the set of outcomes is the sample space, E . All the subsets of the sample space are called events, which are associated with a definite probability of occurrence. The probability associated with any random event is a sum of probabilities of all elementary events that comprise it.

4.1.1 Theory of probability

Although the definition of probability is tricky, the mathematical properties of probabilities are uncontentious and well-defined. The possible outcomes of an experiment are know as random variables x_i with probabilities $P(x_i) = P_i$, then the P_i have to satisfy the axioms of probability which can be taken as

$$P_i \geq 0, \quad P_{i\ or\ j} = P_i + P_j, \quad \sum_i P_i = 1.$$

Moreover, the probability of something not happening must be

$$P_{not\ i} = \bar{P}_i = 1 - P_i = \sum_{i \neq j} P_i$$

If x_i and x_j are two independent events, then the probabilities of x_i AND x_j are multiplied such that

$$p(x_i \cap x_j) = p(x_i)p(x_j).$$

An important concept is that of conditional probabilities, which we denote and is in words the probability that x_i occurs GIVEN that x_j occurs. In general the conditional probability is given by

$$p(x_i|x_j) = p(x_i \cap x_j)p(x_j)$$

We define the probability distribution as the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. There are two quantities which are important properties of any distribution: the expectation value and variance. The expectation value is the long-run average value of the distribution. You can think of it as the average that would be obtained from a very large number of experiments, given the distribution. Since the probability of getting a value x_i is P_i, then for $N \to \infty$ experiments, the number of outcomes equal to x_i would be $N\ P_i$ and so the expectation value is generally given by

$$E(x) = \langle x \rangle = \frac{1}{N} \sum_i x_i (NP_i) = \sum_i x_i P_i \tag{4.1}$$

The variance is the expectation value of the square of the difference of the x_i from the average value

$$V(x) = \langle (x - \langle x \rangle)^2 \rangle \tag{4.2}$$

This can be written in a more useful form

$$V(x) = \langle x^2 \rangle - \langle x \rangle^2 \tag{4.3}$$

It has dimension $[x^2]$ and to get a quantity related to the spread which has dimensions of $[x]$, we take the square root to give the 'root mean square' or RMS. This is also often called the 'standard deviation'.

4.2 Stochastic Processes

4.2.1 Overview

A stochastic model describes a process in which uncertainty is present, i.e. that has some kind of randomness. The word stochastic comes from the Greek word stokhazesthai meaning to aim or guess. In the real word, uncertainty is a part of everyday life, so a stochastic model could represent anything. While for deterministic models the parameter values and the initial conditions fully determine the outcome, stochastic models show their inherent randomness, leading to different outputs from the same set of parameter values and initial conditions.
Focusing on biological models, the dynamics of molecular interactions (such as the binding or dissociation of enzymes) are intrinsically stochastic. Furthermore, molecular systems are often subject to random noise. Intuitively, this means that biomolecular systems can be more accurately modelled as a stochastic process.

4.2.2 The law of large number

In probability theory, the law of large numbers is a theorem that describes the result of performing the same experiment a large number of times. According to the law, the average of the results obtained from a large number of trials can be approximated with the expected value and tend to become closer to the expected value as more trials are performed. Nonetheless, stochastic process are often more computationally expensive to simulate and tend to be more complex to analyze. Deterministic models approximate these processes by applying the law of large numbers. If the number of molecules participating in the system is high enough, the stochastic contributions of each individual molecule compensate and are "averaged out". A common example of this is in nuclear decay. However, where this is not the case (for example, when transcription factors are expressed in low amounts) these approximations tend to result in poorer performance of the model.

4.2.3 Mathematical framework

A stochastic process $X(t)$ can be construed as a series of random variables characterized by their own probability distributions $p(x, t)$. For example, taking $X(t)$ at 2 different time points t_1 and t_2, the probability distribution of each random variable can be denoted as such:

$$P(X(t_1) = x) = p(x, t_1) \tag{4.4}$$

$$P(X(t_2) = x) = p(x, t_2) \tag{4.5}$$

Many stochastic models of biological circuits are assumed to be Markovian, i.e. they follow a Markov process. Markov processes are stochastic processes characterized by its "memory-less" property: the future state of the system is dependent only on the current state of the system, and is independent of its past states. The types of Markov processes depend on the nature of the state space and time parameter: discrete-state discrete-time, discrete-state continuous-time, and continuous-state continuous-time processes as shown in table . These processes can generally be characterized by its state space (all possible states that a system may occupy), transition matrix (a matrix of conditional probabilities or transition rates), and initial distribution of states

Stochastic process	Time	Characteristics	Examples
Branching process	discrete or continuous	population model where each individual's offspring number is drawn from same distribution	colonization of new habitat, spread of new disease
Markov chain	discrete	switches between different states, with probabilities depending on previous state	nucleotide substitutions in DNA sequence evolution
Poisson process	continuous	events that happen independently and with a small probability per unit of time	mutations in lineage of individuals, coalescent process
Wiener process	continuous	random changes in variable but with mean zero	movement of individuals in space

Table 4.1: This table classifies the different stochastic processes

Analyses of stochastic processes either follow the time evolution of the whole probability distribution of the system, or by sampling from the distribution of possible trajectories from the initial state of the system. The latter is achieved by stochastic simulation algorithms.

4.3 The chemical master equation

Now we consider a system which includes molecules of M different chemical species (components) $\{X_1, ..., X_M\}$ that can in total undergo R different irreversible, elementary reactions. These reactions can be of zeroth order (e.g., entry of molecules into an open system), of first order (e.g., degradation of compounds or unimolecular conversion) or of higher order (e.g., dimerization). In the latter case, random encounters of two or more molecules are necessary for the reaction to occur. The j-th reaction can be written as

$$\sum_{i=1}^{M} \beta_{ij} \cdot X_i \rightarrow \sum_{i=1}^{M} \gamma_{ij} \cdot X_i, \tag{4.6}$$

where X$_i$ defines the components in the system and γ_{ij},β_{ij} are the stoichiometric coefficients of the educts and products. We assume that the system is spatially homogeneous so that the copy numbers of the components it contains fully characterize its state. We define the vector $n(t) = (n_1(t), \ldots, n_M(t))^T$, where n_i denotes the copy number of the i-th component, and t is the time variable. In addition, we define the the state-change vector $\nu_j(\nu_{1j}, ..., \nu_{Nj})$, where ν_{ij} is the change in the i-th molecular population caused by one R_j reaction, so if the system is in state x and one R_j reaction occurs, the system immediately jumps to state $X + \nu_j$

In the CME framework, Markov property holds since the system state is modeled as a continuous-time stochastic process. Then, the probability distribution of future system states only depends on the present state, but not on past states (memorylessness). In addition, the state space is discrete, following the above definition. The CME is formulated as:

$$\dot{p}_n = \sum_{m=1}^{M} (\pi(n,m)p_m(t) - \pi(m,n)p_n(t)), \tag{4.7}$$

where $p_n(t)$ is the probability of being in state n at time t, $\dot{p}_n$ its time derivative and $\pi(n,m)$ is the probability per infinitesimal time unit of a transition from m to n. The temporal evolution of p_n is determined by the balance between transitions leading to state n and transitions away from n. Now, we want to define π, taking into account the reaction system in 4.6. Single reactions control transitions considering infinitesimal intervals. All possible transitions between states can be captured by the stoichiometric matrix A with entries $a_{ij} = \gamma_{ij} - \beta_{ij}$ and columns $a_j = (a_{1j}, \ldots, a_{Mj})^T$,so that the CME is expressed as

$$\dot{p}_n = \sum_{j=1}^{R} (w_j(n - a_j) \cdot p_{n-a_j}(t) - w_j(n) \cdot p_n(t)), \tag{4.8}$$

where $w_j(n)$ is the probability per infinitesimal time unit for the j-th reaction to occur, when the system is in state n (the propensity of the j-th reaction). The propensities can be formulated as

$$w_j(n) = k_j \prod_{i=1}^{M} \binom{n_i}{\beta_{ij}} \tag{4.9}$$

where k_j is the stochastic reaction constant. It results from physical properties of the reaction (such as activation energy, complexity) and by environmental conditions like temperature. The propensities highlight the combinatorial probability of random encounters of the educts, taking into account reactive collisions of the components, where β_{ij} out of n_i molecules of the i-th component are involved. A close inspection reveals that the CME is actually a set of coupled ODEs, with one equation for every possible combination of reactant molecules. It is therefore not surprising that the CME can be solved analytically for only a few simple cases, and even numerical solutions are prohibitively difficult in other cases.

4.4 Chaos Theory

The line between the two modelling approaches deterministic and stochastic is further blurred by the development of chaos theory. Chaos theory involves deterministic models that can have different outcomes with slight changes in the model. Some argue that most stochastic models are in fact chaotic deterministic models, a thought which is summed up nicely by Lothar Breuer of the University of Kent:
"A mountain stream, a beating heart, a smallpox epidemic, and a column of rising smoke are all examples of dynamic phenomena that sometimes seem to behave randomly. In actuality, such processes exhibit a special order that scientists and engineers are only just beginning to understand. This special order is deterministic chaos', or chaos, for short."
And while chaos theory supplies a viable alternative to deterministic or stochastic models, it's applications to probability theory is still in its infancy.

Chapter 5

Stochastic Model Examples

5.1 Gillespie Algorithm

The Gillespie stochastic simulation algorithm is an event-based algorithm that randomly generates the next event state and time using the probability distribution specified by the transition matrix. Because the CME can rarely be solved for the probability density function of X(t), perhaps we should look for a way to construct numerical realizations of X(t), i.e., simulated trajectories of X(t) versus t. This is not the same as solving the CME numerically, as that would give us the probability density function of X(t) instead of a random sample of X(t). The key to generating simulated trajectories of X(t) is not the function p(x,t). The approach of Gillespie has as a central notion the following definition of reaction probability density function: $P(\tau, j)$ is the probability that, given the state $X = (X_1, ..., X_M)$ at time t, the next reaction will occur in the infinitesimal time interval $(t + \tau, t + \tau + dt)$, and will be an R_j reaction. Formally, this function is the joint probability density function of the two random variables time to the next reaction (τ) and index of the next reaction (j), given that the system is currently in state X. Given a set of R reactions $\{R_1, ..., R_R\}$ and a current time t, we draw two random numbers r_1 and r_2 from the uniform distribution in the unit interval in order to determine the value of

$$\tau = \frac{1}{\sum_{\nu=1}^{R} w_\nu} ln\Big(\frac{1}{r_1}\Big)$$

$$\text{and} \quad j = \text{the smallest integer satisfying } \sum_{j'=1}^{j} w_{j'} > r_2 \sum_{\nu=1}^{R} w_\nu$$

Hence, we have the following stochastic simulation algorithm (SSA) for constructing an exact numerical realization of the process X(t):
1. The initialization happens as the number of molecules in the system, reaction constants, and random number generators is chosen.
2. The time t+τ at which the next reaction will occur is randomly chosen with τ exponentially distributed with parameter $\sum_{\nu=1}^{R} w_\nu$;
3. The reaction R_{j} that has to occur at time t+τ is randomly chosen with probability w_j dt.
4. At each step t is incremented by τ and the chemical solution is updated. Hence, at each step the probability density function is

$$P(\tau, j) = \underbrace{e^{-\sum_{\nu=1}^{R} w_\nu \tau}}_{\text{time elapsed}} \quad \underbrace{w_j \; dt}_{\text{next reaction}} \tag{5.1}$$

and gives the probability that the next reaction will occur in the time interval $(t + \tau, t + \tau + dt)$ and will be R_j.
5. Then, the process is iterated, going back to Step 1 unless the number of reactants is zero or the simulation time has been exceeded.

5.2 Tau leaping

The Gillespie algorithm is computationally expensive and so many modifications and adaptations exist, such as the tau-leaping algorithm. Tau-leaping was developed by Gillespie to increase the computational speed of the SSA, which is an exact method. Instead of computing the time to every reaction, this algorithm approximates the process and attempts to leap in time, executing a large number of reactions in a period τ. This algorithm is computationally faster; however, the approximation removes the "exact" connection to the solution of the chemical master equation method for the system. With the system in state X at time t, let us suppose there exists a $\tau > 0$ that satisfies the following leap condition: During $[t, t+\tau)$, no propensity function is likely to change its value by a significant amount. With w_j remaining essentially constant during $[t, t+\tau)$, it then follows from the fundamental premise that the number of times reaction channel R_j fires in $[t, t+\tau)$ is a Poisson random variable with mean (and variance) $w_j\tau$. Therefore, to the degree that the leap condition is satisfied, we can approximately leap the system ahead by a time τ by taking

$$X(t+\tau) = X(t) + \sum_{j=1}^{R} \mathcal{P}_j(w_j\tau)\nu_j \tag{5.2}$$

where $\mathcal{P}_j(m_j)$ is a statistically independent Poisson random variable with mean (and variance) m_j. Example

Chapter 6

Additional Modelling Frameworks

6.1 Compartmentalisation

Compartmentalisation is a modelling technique used to replicate the different compartments of a given system. For synthetic biology, this translates to cells, with varying levels of permeability between compartments. Additionally, you could assume that a volume with a low viscosity is made up of a number of compartments, so as to simplify the concentration distribution to discrete values for each compartment instead of a complicated, continuous variable.

Consider this example, where we have two compartments A and B with a partially permeable membrane between them. The molecules are created in component A, and diffuse into component B. We can use Fick's Law to model this process. Usually concentration is a continuous variable in a volume. If there is symmetry in 2 dimensions, concentration can be simplified to a continuous variable over a length instead. With the membrane, this can become quite complex, as there is a discontinuity at the membrane.

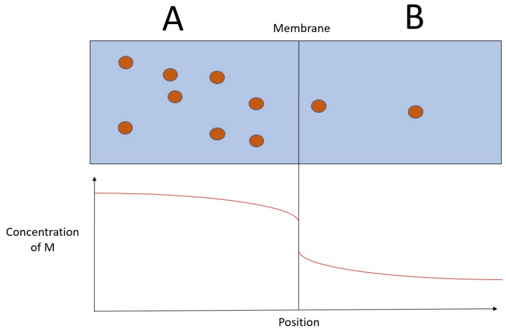

Figure 6.1: Two compartments A and B, one with a greater concentration average than the other and a permeable membrane between the two. Concentrations are continuous

Instead, again for the sake of simplicity for creating an initial model, we assume the concentration is uniform throughout A and throughout B (see Fig 2). In this way, we can approximate the continuous concentration variation to be discrete: concentration in A and concentration in B. We don't need to assume the volumes are the same, since we are handling concentrations.

Given these approximations, we can use a simplified version of Fick's Law. We define the molecule as M, and the concentration in A as *[M]*$_A$ and in B as *[M]*$_B$

$$\frac{d[M]_\mathrm{A}}{dt} = -\frac{DS}{L}[M]_\mathrm{A} + \frac{DS}{L}[M]_\mathrm{B} = -\frac{DS}{L}([M]_\mathrm{A} - [M]_\mathrm{B}) \tag{6.1}$$

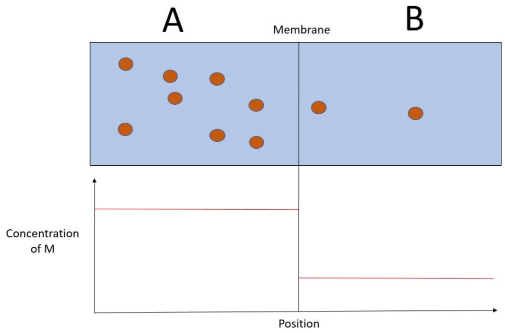

Figure 6.2: Two compartments A and B, one with a greater concentration than the other and a permeable membrane between the two. Concentrations are discrete and uniform in A and in B

and conversely

$$\frac{d[M]_{\mathrm{B}}}{dt} = \frac{DS}{L}([M]_{\mathrm{A}} - [M]_{\mathrm{B}}), \tag{6.2}$$

where D is the Diffusion Coefficient with dimensions $[Length^{-1}Time^{-1}] = [L^{-1}T^{-1}]$, S is the surface area of the membrane $[L^2]$ and L is the thickness of the membrane $[L]$. For a given system, D, S and L can all be integrated into one constant. However, D will vary between molecules. This is essentially how easy it is for a given molecule to fit through the membrane. This can rather crudely be defined as the size of the molecule, although some molecules may be too big to fit at all, or may be so small the membrane is irrelevant to slowing down diffusion. The shape of the molecule will also affect diffusion rate.

The most thorough way to determine the constants would be through empirical data. It would also be better to incorporate the continuous model, where diffusion through the membrane is a stochastic process, but this is certainly a good starting point. It is also possible to approximate the constants for the sake of the simulation, just to observe how the compartmentalisation affects the system.

6.2 Whole cell models

Whole cell models are a particular type of cell model where the entire dynamics of the cell are taken into account, such that no assumptions about cell dynamics are made. In other words, the phenotype can be accurately predicated from genotype. Instead of considering a novel system and needing data to develop a bespoke model, a whole cell model is a generalised model, allowing changes in the cell, such as introduction of an engineered plasmid, to be easily incorporated and results easily obtained. These incorporate everything from thermodynamics to molecular interactions to compartmentalisation of the cell to the extracellular environment.

However, due to the complexity of the cell, these models are extremely difficult to develop. They require extensive empirical data to inform the design of the model, then integration of different models into one whole structure, then extensive validation. As of when Version 1 of this package was developed [October 2020], whole cell models are limited in their extent even in simple cells. Click *here* to read more.

Chapter 7

Mathematical Modelling in Practise

Mathematical modelling is a very powerful tool in the scientists toolbox. It allows us to quickly, easily and cost-effectively obtain lots of data without need for experimentation. They provide a simplified version of reality, accounting for only the essential features of a system and ignoring the unnecessary details.

We know objects fall to earth thanks to the classical physics models sir Isaac Newton developed, so we don't need to test if new objects fall to earth due to gravity. However, Einstein found there were limitations to these models, from which he developed the more generalised theories of relativity.

The same is true in cellular biology, but to a much greater extent. Our knowledge of the dynamics of a cell is extremely limited currently, and will inevitably be built upon. For example, in the basic gene expression model developed in 3.2, we made a number of assumptions: processes are continuous, not discrete; promoters and RBS are arbitrarily strong and the transcription factor, ribosomes and tRNAs were uniformly, consistently and infinitely dense; the additional gene causes no strain on the cell; the protein expressed isn't toxic to the cell; mutation rate is zero; and the list goes on. There are certainly more complex models already developed which account for some of these, but even these aren't complete, and assumptions are still made.

As good scientists, we must therefore be aware of the extent to our models are valid. We don't have a complete whole cell model yet, so when novel systems are created, they must have some empirical validation. Given this, there are two main frameworks for using a model. It can either be predictive, where the model influences the design of a project, or it can be exploratory, where the empirical data helps to build the model.

7.1 Exploratory Modelling

Exploratory modelling is a framework for developing a mathematical model when very little knowledge of dynamics is known, appreciating the limitations of the model and requiring empirical data to validate and develop the model. The empirical data influences the model design. In general, the following steps are taken:

1. A basic model is determined before any experimentation based on prior knowledge obtained from previous reading and research.
2. The model is empirically tested to determine strengths and weaknesses.
3. The model is developed based on this data, adding improvements and modifications.
4. Steps 2 and 3 are repeated to refine the model until the necessary level of accuracy is met, based on the requirements of the model.

At its core, this type of modelling is a form of educated guesswork. It is therefore long, arduous and costly. However, given there is such limited knowledge of cell dynamics, this framework ensures the model is well validated at every step.

7.2 Predictive Modelling

Predictive modelling is a framework where behaviour of a system is predicted based on the model, without empirical data to support that prediction. In other words, we rely on the accuracy and validity of the model to predict unknown behaviour. This can be based on an exploratory model developed for a specific system, or based on previous research done around the field.

While in an ideal world, there is no better data than empirical data, a model has many benefits over carrying out extensive experimentation. It is cheaper, faster, repeatable and has thorough logging (that is, the code can be read and errors logged, whereas individual pipetting needs to be logged by hand) allowing for mistakes to be easily found and rectified. This can be invaluable for any field of science, synthetic biology included.

For example, in combinatorial design, the design space (the total number of iterations) can be enormous, where throughput and time required are unrealistic. Even for a small design space, perhaps a certain reagent or material is very expensive, and so reducing it even a little bit would save a lot of money. Perhaps an experiment would take months to finish, meaning we are banking on the system working and committing months to show this. A model can give some indication of whether a system might work, and even how to change it if it doesn't.

Chapter 8

Finishing Statement

In this summary, we have covered all the necessary information to provide a jumping board to approach mathematical modelling for synthetic biology. We covered the theory of several essential models, and elaborated on how we could develop the model further with additional complexity. This gives the necessary foundation upon which further reading can be conducted for more specific or specialised models and scenarios.

Some additional general modelling resources have been provided below. We also encourage readers to find specific examples for their given system as well.

Many thanks to Professor Guy-Bart Stan and the following IGEM 2020 teams for their input and feedback on this guide: Hamburg, Gaston_Day_School, Korea_HS, CLS_CLSG_UK and UNILausanne.

Bibliography

These papers provide further reading:

Chandran D, Copeland WB, Sleight SC, Sauro HM. Mathematical modeling and synthetic biology. Drug Discovery today. Disease Models. 2008 ;5(4):299-309. doi: 10.1016/j.ddmod.2009.07.002

Karr JR, et al. A whole-cell computational model predicts phenotype from genotype. Cell. 2012;150:389–401. doi: 10.1016/j.cell.2012.05.044

Purcell O., Jain B., Karr J. R., Covert M. W., Lu T. K. (2013). Towards a whole-cell modeling approach for synthetic biology. Chaos 23:025112 doi: 10.1063/1.4811182

Zheng, Y., Sriram, G. Mathematical modeling: bridging the gap between concept and realization in synthetic biology. Journal of biomedicine & biotechnology, 541609 (2010)

Guido, N., Wang, X., Adalsteinsson, D. et al. A bottom-up approach to gene regulation. Nature 439, 856–860 (2006). doi: 10.1038/nature04473

Winstead, C., Madsen, C., and Myers,C. iSSA: An incremental stochastic simulation algorithm for genetic circuits. Proceedings of 2010 IEEE International Symposium on Circuits and Systems, Paris. 553-556 (2010), doi: 10.1109/ISCAS.2010.5537539.

Neupane, T., Zhang, Z., Madsen, C., Zheng, H., Myers, C. Approximation Techniques for Stochastic Analysis of Biological Systems. 327-348 (2019). doi: /10.1007/978-3-030-17297-8_12