Synthetic Biology for Mathematicians and Engineers

iGEM UWaterloo Mathematics and Modelling Training Document
Michael Astwood, University of Waterloo
mrastwoo@uwaterloo.ca

February 2020

Contents

1 Introduction to Cellular Biology

1.1 Prokaryotic Cell Structure

All living things are made of cells. Cells are small capsules that perform chemical reactions and store information. The basic structure of a cell is as follows:

1. Outer membrane made of oily molecules (called phospholipids)
2. Inner fluid called cytoplasm
3. Inner 'skeleton'-like structure called the cytoskeleton which transports molecules

There are two large classes of cells, which tend to exhibit very different properties. These are called eukaryotes and prokaryotes. Eukaryotes have structures inside of them called organelles, and also store their DNA in a nucleus.

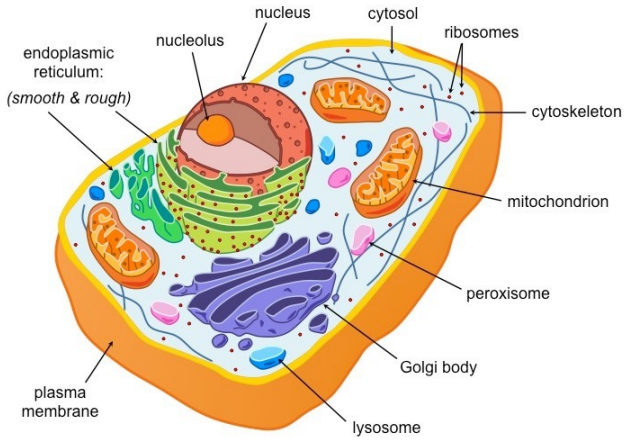

Figure 1: Simplified diagram of a eukaryote.

Since a eukaryote stores its' DNA in the nucleus, it can be difficult to introduce new DNA to the cell. In an iGEM project we are typically looking to insert DNA into a cell in order to provide functionality, but it can be hard to get a eukaryote to actually use the DNA we give it. This is because not all of the DNA in a cell is used at a given moment, and we must ensure that the DNA is being expressed! Let's focus on prokaryotes instead.

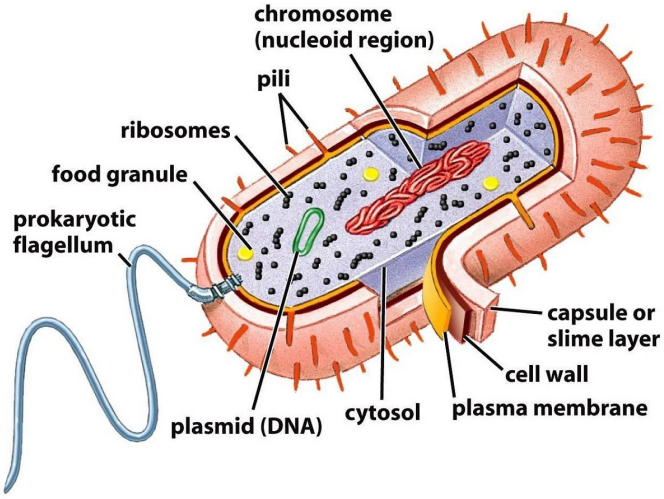

Figure 2: Simplified diagram of a prokaryote.

Let's break down what's going on in figure 2. Here we see the following structures (in clockwise order):

1. Capsule or Slime layer: This outer layer of the cell protects the inside of the cell, and has proteins and other molecules embedded inside of it in order to perform functions outside of the cell.
2. Cell Wall: This layer of the cell is made up of a substance called peptidoglycan, which ensures the rigidity and safety of the internal structure.
3. Plasma Membrane: This membrane on the inside of the cell wall consists of phospholipids, which are oily molecules that stick together in order to form a semi-permeable membrane.
4. Cytosol: This is the liquid which composes the majority of the cytoplasm.
5. Plasmid: Plasmids are circular strands of DNA which float in the cytoplasm.
6. Flagellum: A flagellum is a long fibre composed of a protein called flagellin. The flagellum spins or waves in order to move the cell around. It can also perform sensory functions for the cell.
7. Food Granule: When small pieces of food enter the cell, they can either be wrapped in a phospholipid membrane layer or just float around. These pieces of food are called granules.
8. Ribosome: The ribosome is a large system of molecules which produces proteins for the cell in a process known as mRNA translation.
9. Pilus: Pili are small fibres on the outside of a cell which perform a function known as conjugation, where two cells will transfer DNA to one another.
10. Chromosome: The nucleoid region is where most of the DNA in the cell is stored. This coagulation of DNA is called a chromosome.

We will see many of these systems in much greater depth later.

1.2 Common Molecules Within a Cell

There are a few major classes of molecules which make up the systems in a cell. They are:

1. Deoxyribonucleic Acid (DNA). DNA is composed of nucleotides, which are small molecules which join together along a common 'backbone' made up of phosphorus and carbon. These nucleotides form a linear sequence of molecules which may be labelled according to the names of the nucleotides: Adenine, Guanine, Cytosine, and Thymine (AGCT). These chains bind together into a ribbon-like pattern, and form what is known as a double-helix structure. This is seen in figure 3
2. Ribonucleic Acid (RNA). RNA is a molecule with very similar structure to DNA, except it is composed of Uracil instead of Thymine. RNA only forms a single stranded structure, though the nucleotides may bind together to form small "hairpin" shaped structures.
3. Amino Acids. These are small molecules which combine together in order to form proteins. There are 21 common amino acids, and they each have particular properties that can impact the structure of a protein greatly.
4. Protein. These are large structures which are formed out of one or more chains of amino acids joined by covalent bonds. These structures are generated by a ribosome, by matching sequences of RNA nucleotides to particular amino acids.
5. Lipids. This is a class of molecule which includes fats, many oils, and other molecules. Lipids are generally composed of long chains of carbon, which may branch or bind together.

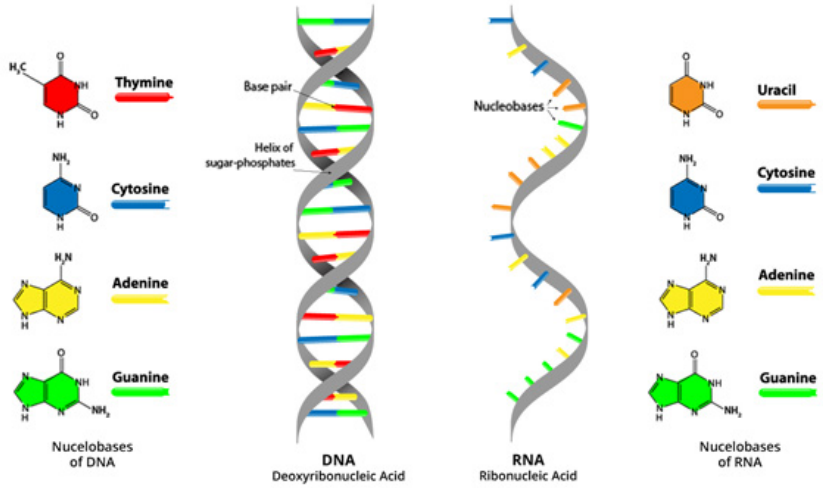

Figure 3: Simplified diagram of DNA and RNA.

1.3 Processes of Gene Expression

How exactly does DNA get turned into instructions for the cell? There are a few molecules we must know about which are going to be important here:

1. mRNA (Messenger RNA): These are long strands of RNA which are read letter-by-letter and used to generate proteins.
2. tRNA (Transfer RNA): These are T-shaped strands of RNA which bind to amino acids. These enter the ribosome and bind to the mRNA strand in the translation step.
3. rRNA (Ribosomal RNA): In fact, the ribosome itself is made up of RNA! This is because RNA is a simple molecule that is easy to produce, and it can form a variety of structures.

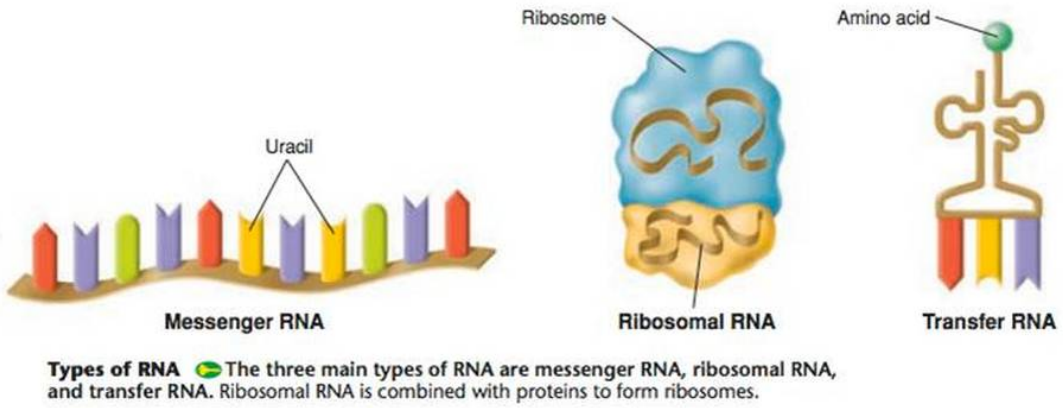

Figure 4: Simplified diagram of different types of RNA.

1.3.1 Transcription

The DNA in a cell is joined together by weak bonds called hydrogen bonds. When the cell is ready to produce something from the DNA, there will be an increased presence of a protein called RNA Polymerase, which generates strands of RNA.

This protein will find its way to the DNA and split open these hydrogen bonds, creating a "zipper" effect.

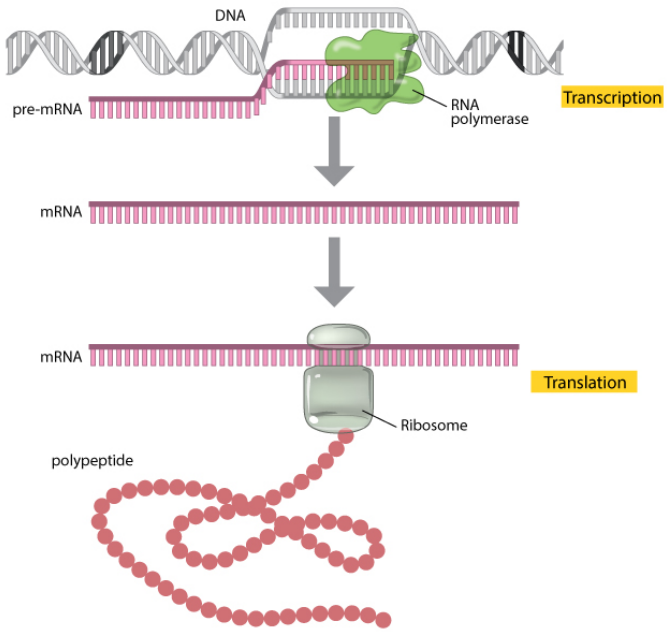

Figure 5: Simplified diagram of transcription

The RNA polymerase reads the DNA three letters (ATGC) at a time. Each sequence of three letters is called a *codon*. When the RNA Polymerase comes across the "promoter" region, it will begin to bind nucleotides together opposite the strand of DNA. This produces a long strand of RNA coming out the other end as seen in figure 5. Finally, the RNA Polymerase will come across the terminator and the RNA strand will be released and the DNA will be allowed to settle back into its initial state.

As the RNA leaves the polymerase, it is termed "messenger RNA", because it is delivering the information contained in the DNA somewhere else. We will refer to this as mRNA.

In order to be turned into a protein, this mRNA enters a ribosome, and begins the process of translation.

1.3.2 Translation

When mRNA comes in contact with a ribosome, it begins a process called transcription.

1. First, the ribosome's two substructures clamp around the mRNA strand. This process is called initiation.
2. A Methionine tRNA is present at the main site on the ribosome
3. The ribosome slides the mRNA along until it comes across the START codon which codes for methionine
4. tRNA molecules then enter the ribosome
5. The first codon on the mRNA strand is matched up with the tRNA's primary codon (seen in figure 4).
6. The amino acid on the matching tRNA molecule is bound to the ribosome.
7. A second tRNA molecule enters, and the mRNA is shifted to the next codon.
8. If this is a match, amino acid number 2 will bond with the first, beginning a short chain of amino acids.
9. This process continues until the ribosome reaches the STOP codon, and then releases the long chain of amino acids and mRNA into the cell.
10. The mRNA may come in contact with another ribosome. The process repeats.

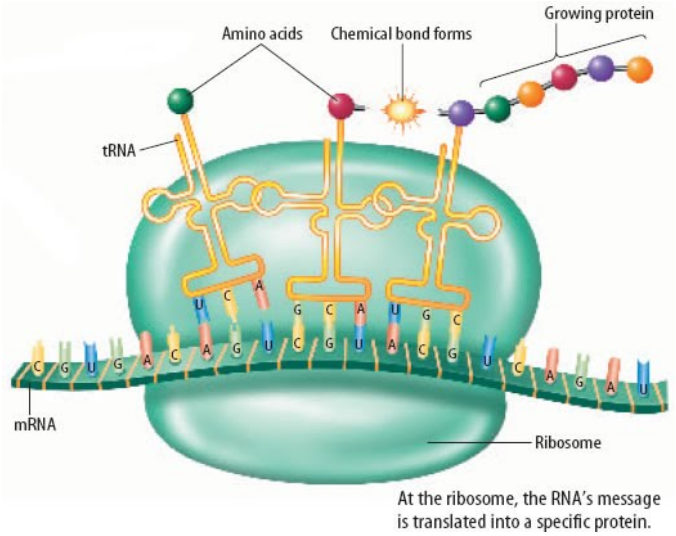

Figure 6: Simplified diagram of translation.

1.4 Engineering Gene Expression

We already heard that not all of the DNA in a cell is expressed at a given time, so why does this happen? There must be some factors which affect the ability of RNA polymerase to bind to the DNA at a specific site. These sites are called Promoters, and can be activated or deactivated.

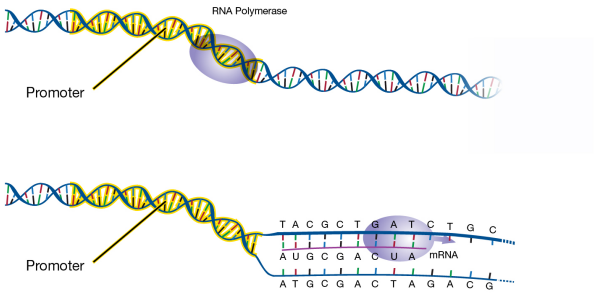

Figure 7: Simplified diagram of a promoter site.

These promoters may be more or less likely to bind to RNA Polymerase. This is due to a number of reasons, including the fact that it might take too much energy for an RNA Polymerase to bind to the promoter. This is important in cellular biology - it takes energy to do anything, and if that process requires too much energy it won't happen as often. In prokaryotes, a promoter typically consists of two short sequences of nucleotides located 10 or so positions upstream from the sequence of interest, along with other sequences nearby. These short sequences may mutate, and could have a mismatch with the RNA Polymerase - this might not completely stop expression, but it would take longer for a polymerase to find the right spot.

In order to do synthetic biology, we typically choose a promoter to go along with our gene of interest. This promoter is chosen based on whether the literature tells us it's strong or weak - a promoter which is typically slow is called weak, and vice versa. Note that this isn't exact, and that the conditions under which a promoter may work or not will vary depending on more complicated factors.

Many genes may be placed under the same promoter, causing the cell to produce a set of proteins rather than one at a time. Such a structure is called an operon.

In an engineering setting, it would be desirable to find molecules which can increase or decrease the level of gene expression in a cell. Such a molecule is called a gene regulator, and there are a few major types:

1. Repressors: These regulatory molecules are proteins which bind to DNA adjacent the genes which they control. These molecules slow down the RNA polymerase from translating a particular gene.
2. Activators: These regulatory molecules are proteins which bind to the promoter, increasing the level of transcription.
3. Inducers: These molecules are typically small compounds which can either activate or repress gene expression by interacting with an activator or repressor.

Interesting systems called genetic circuits (or gene regulatory networks) may be built by considering the interactions between these regulatory molecules. We will explore this in depth in the modelling section.

2 Intro to Mathematical Modelling with Differential Equations

This first section on mathematics will be a short overview of differential equations, their behaviour, and how to solve them. I'll begin by defining a differential equation. First let me show you a few examples:

$$\frac{d^2y}{dx^2} + 4y = 0 \implies y = c_1\sin(2x) + c_2\cos(2x) \tag{1}$$

$$\frac{dy}{dx} + 2y = 0 \implies y = ce^{-2x} \tag{2}$$

$$\frac{dy}{dx} = e^{-y}(2x-4) \implies y = \ln\left(x^2 - 4x - 4\right) \tag{3}$$

Definition 1: Ordinary Differential Equation

An ordinary differential equation in y is an equation relating a function $y(x)$ to its' derivatives with respect to x:

$$G(y^{(n)}, y^{(n-1)}, ..., y, x) = 0 \tag{4}$$

This equation must be satisfied for all x.

The *order* of a differential equation is the order of the highest derivative in the equation. For example, consider the equation:

$$G(y'', y', y) = \frac{d^2y}{dx^2} + 3\frac{dy}{dx} + \sin(y) + 4 = 0$$

This is a second order equation, since the highest derivative it contains is a second derivative. There are a few different classes of ordinary differential equation, which are important to remember. The key class is that of linear ordinary differential equations, which are very well understood.

Definition 2: Linear Ordinary Differential Equation

An ordinary differential equation in $y(x)$ is called *linear* if the equation $G = 0$ has the following form:

$$A_n y^{(n)} + ... + A_1 y' + A_0 y = F \tag{5}$$

Where each A_i may depend on x, as can the right hand side F.

The solutions to linear differential equations are well behaved, and these equations can be solved with great precision. Here are some examples of linear equations:

$$2y'' + 3y' + 2y = \sin(x)$$
$$\cos(x)y'' + 3y = 0$$
$$x^2y'' + xy' + y = 5$$

The key thing here is to note that y only appears in its' derivatives - there are no functions which are composed with y. Getting rid of composition makes the equation a lot easier to solve. Using Taylor series, it is often possible to linearize a differential equation so that you can approximate the solution in some region.

3 Chemical Kinetics

Now we will start to use these differential equations in order to do some modeling relevant to iGEM! This chapter will be Let's consider the problem of a chemical reaction network. You have some input reactants and their concentrations, and you would like to see how the concentration of the products will change over time. There are a number of questions you can ask about the reaction.

1. Will the concentrations equilibriate?
2. What are the concentrations at equilibrium?
3. What is the effect of a catalyst on the reaction rate?
4. Is the reaction rate appreciable?
5. Are there reactions in the network that slow down the system?
6. What if someone manually introduces new reactants?

So how can we use differential equations to answer these questions? The link is called the Law of Mass Transfer:

Theorem 1: Law of Mass Transfer

The rate of a chemical reaction is proportional to the product of the concentrations of the reactants. Suppose $A_1 + ... + A_n \to B$ is a chemical reaction. Then the concentration of the reactant A_i follows the differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}[A_i] = -k[A_1]...[A_n]$$

This gives us n first order differential equations which we must solve.

This law may be derived from some statistical mechanics. It essentially states that the probability of a reaction happening is proportional to the probability of a collision between the reactants. The big assumption here is that there are a lot of molecules! Let's consider a few examples.

3.1 Example 1: Reversible Production and Decay

This first example represents the reversible conversion of a reactant A into a product B, which is then removed from the system at a rate k_{decay}.

$$\mathrm{A} \underset{k_-}{\overset{k_+}{\rightleftharpoons}} \mathrm{B} \xrightarrow{k_{decay}}$$

From the law of mass transfer, we can find the following relations:

$$\frac{\mathrm{d}}{\mathrm{d}t}[A] = -k_+[A] + k_-[B]$$
$$\frac{\mathrm{d}}{\mathrm{d}t}[B] = k_+[A] - k_{decay}[B] - k_-[B]$$

3.1.1 Solution of system

Now, this is a system of differential equations which we must solve. Let's relabel these concentrations to make them easy to write. Let $a(t) = [A]$, and $b(t) = [B]$. Then we have:

$$\frac{\mathrm{d}a}{\mathrm{d}t} = k_- b - k_+ a$$
$$\frac{\mathrm{d}b}{\mathrm{d}t} = k_+ a - k_{decay} b - k_- b$$

Which can be written in matrix form:

$$\begin{bmatrix} a' \\ b' \end{bmatrix} = \begin{bmatrix} -k_+ & k_- \\ k_+ & -k_{decay} - k_- \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

Writing the system in matrix form makes the differential equation a lot easier to solve. I will give you the formal way to solve it, and then show you how to do it in practice. So:

$$\begin{bmatrix} a' \\ b' \end{bmatrix} = \begin{bmatrix} -k_+ & k_- \\ k_+ & -k_{decay} - k_- \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

Can be rewritten in vector form as:

$$\vec{x}' = M\vec{x}$$

Where M is called the rate matrix. This is easiest to solve if we can find combinations of a and b so that $M\vec{y} = \lambda\vec{y}$ for some number m. You might recall that these special pairs $\vec{y}$ are called eigenvectors, and the values of λ are called eigenvalues. The way we find the eigenvectors and eigenvalues is called diagonalization, and the process goes like this:

$$\begin{aligned} M\vec{x} - \lambda\vec{x} &= 0 \\ (M - \lambda\,\mathbb{1})\vec{x} &= 0 \\ \implies \det(M - \lambda\,\mathbb{1}) &= 0 \end{aligned}$$

The determinant $\det(M - \lambda\,\mathbb{1})$ is a polynomial in λ, and it is called the characteristic polynomial of M. The polynomial has the form:

$$\det(M - \lambda\,\mathbb{1}) = (-k_+ - \lambda)(-k_{decay} - k_- - \lambda) - k_- k_+ = 0$$

And the zeros are:

$$\lambda = \frac{1}{2}\left[\pm\sqrt{k_+^2 + 4k_+k_- - 2k_+(k_{decay} + k_-) + (k_{decay} + k_-)^2} - k_+ - (k_{decay} + k_-)\right]$$

Suppose that the eigenvalues are different (the descriminant is nonzero). Then there are two independent eigenvectors of M, and we can write the solution to the differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}\vec{y} = \lambda\vec{y} \implies \vec{y} = e^{\lambda t}\vec{y}$$

Let $\vec{y}_1$ and $\vec{y}_2$ be the two eigenvectors of M. Then the solution to the differential equation is:

$$\vec{x} = C_1 e^{\lambda_1 t}\vec{y}_1 + C_2 e^{\lambda_2 t}\vec{y}_2$$

Case 0: If both the eigenvalues are zero, then the system is completely stationary.

Case 1: Let's suppose that the eigenvalues are both real and have opposite sign. Let's also assume that $\lambda_1 \neq \lambda_2$. One is negative and one is positive. This would mean that as time goes on, one of the solutions diminishes and we are left with a single combination $\vec{y}_i$ of $[A]$ and $[B]$ growing exponentially large. That is: either $[A]$ or $[B]$ grows forever! This can happen if for example, k_+ is very very large and the other rate constants are very very small.

Case 2: Let's suppose that the eigenvalues are both real and have the same sign. Let's also assume that $\lambda_1 \neq \lambda_2$. Then the trace of the matrix determines the stability of the system. If $-(k_+ + k_{decay})$ is positive, then the exponents are going to blow up! We will have very large concentrations as time goes on - this is not desired, because it indicates buildup rather than production! However, if $-(k_+ + k_{decay})$ is negative, then the exponents will shrink and we will have no buildup.

Case 3: The final case is when the eigenvalues are both complex. Then the same thing applies to the trace. If $-(k_+ + k_{decay})$ is positive, then the exponents are going to blow up. We will have very large concentrations as time goes on - this is not desired, because it indicates buildup rather than production! However, if $-(k_+ + k_{decay})$ is negative, then the exponents will shrink and we will have no buildup. The only thing different about the eigenvalues being complex is that the concentrations of A and B are going to oscillate.

In practice, case 3 should not be significant for a chemical reaction. However, this is a very rich example for the study of chemical kinetics, so let's look at a few other features of the system.

3.1.2 Rapid Equilibrium Approximation

This reaction is the prime example to use for something called the rapid equilibrium approximation. Suppose that this reaction was just one chunk of a larger system which perhaps uses B to produce something new. It is often useful to make approximations to the system, particularly when we want to examine limiting behaviour or when we want to make the computations less heavy.

Recall that the reaction has two steps. First, A gets converted into B (or vice versa!). Then B decays or gets converted into something else.

The rapid equilibrium assumption is that the first step happens much more quickly than the second step. It takes much less time for A and B to turn into each other than it does for B to decay. This would mean that the reaction $A \iff B$ will equilibriate much faster than B decays.

When $A \iff B$ is in equilibrium, we have:

$$\frac{\mathrm{d}a}{\mathrm{d}t} = 0 \implies \frac{[A]}{[B]} = \frac{k_-}{k_+}$$

However, the rapid equilibrium approximation will assume that this state is achieved almost immediately, so that we may instead focus on how much B is left in the system.

We may then apply this approximation to the second step! Let's define a new variable $c(t) = a(t) + b(t)$. Then we get:

$$c(t) \approx \frac{k_- + k_+}{k_-} a(t), \qquad c(t) \approx \frac{k_- + k_+}{k_+} b(t)$$

Since by the rapid equilibrium assumption, the concentration of A does not change, this turns the second differential equation into the following.

$$\frac{\mathrm{d}c}{\mathrm{d}t} = -k_{decay} b(t) = \frac{k_{decay} k_+}{k_+ + k_-} c(t)$$

Essentially what we have done is reduce the system from $[A \iff B \to D]$ to $[(A + B) \to D]$. This approximation does not work in all cases! In fact, it will break down whenever A is changing or being introduced - and this happens often! The solution to this problem is to instead use the quasi-steady-state approximation. You can see this in Brian's book.