MCB137L/237L: Physical Biology of the Cell
Spring 2025
Homework 5
(Due 2/25/25 at 2:00pm)

Hernan G. Garcia

1 Dynamics of A $\rightarrow$ B reactions.

One of the most interesting topics in science is how we have learned to probe deep time. Surprisingly, DNA sequence has permitted us to explore deep time in the biological setting. Of course, biology and the dynamics of the Earth are not independent phenomena and the point of the rest of this problem is to better understand the details of how scientists figure out how old the Earth is as well as how old various fossil-bearing strata are. To that end, we will first consider a simple model of the radioactive decay process for potassium-argon dating methods, recognizing that there are many other dating methods that complement the one considered here.

Potassium-Argon dating

Potassium-argon dating is based upon the decay of ^{40}K into ^{40}Ar. To a first approximation, this method can be thought of as a simple stopwatch in which at t = 0 (i.e. when the rocks crystallize), the amount of ^{40}Ar is zero, since it is presumed that all of the inert argon has escaped. We can write an equation for the number of potassium nuclei at time $t + \Delta t$ as

$$N_{\mathrm{K}}(t + \Delta t) = N_{\mathrm{K}}(t) - (\lambda \Delta t) N_{\mathrm{K}}(t). \tag{1}$$

Stated simply, this means that in every small time increment Δt, every nucleus has a probability $\lambda \Delta t$ of decaying, where λ is the decay rate of ^{40}K into ^{40}Ar. We also employ the important constraint that the number of total nuclei in the system must remain constant, so that

$$N_{\mathrm{K}}(0) = N_{\mathrm{K}}(t) + N_{\mathrm{Ar}}(t), \tag{2}$$

where $N_{\mathrm{K}}(0)$ is the number of ^{40}K nuclei present when the rock is formed, $N_{\mathrm{K}}(t)$ is the number of ^{40}K nuclei present in the rock at time t, and $N_{\mathrm{Ar}}(t)$ is likewise the number of ^{40}Ar nuclei present in the rock at time t. In this part of the problem you will use equations 1 and 2 to construct differential equations to find the relationship between $N_{\mathrm{K}}(t)$, $N_{\mathrm{Ar}}(t)$, and t.

(a) Using equations 1 and 2 as a guide, write differential equations for $N_{\mathrm{K}}(t)$ and $N_{\mathrm{Ar}}(t)$. How do these two expressions relate to one another?

(b) Next, we note that the solution for a linear differential equation of the form $\frac{dx}{dt} = kx$ is given by $x(t) = x(0)e^{kt}$. Use this result to solve for $N_{\mathrm{K}}(t)$.

(c) Use the constraint encapsulated by equation 2 to write an equation for the lifetime of the rock, t, in terms of the ratio $\frac{N_{\mathrm{Ar}}}{N_{\mathrm{K}}}$.

Age of the Galapagos Islands

The potassium-argon dating method described above has been used in several contexts central to some of the most important evolutionary questions in biology. As we go from West to East in the Galapagos Archipelago, the ages of the islands increase, with Santa Cruz older than Isabella, for example. But how are these numbers known and what evidence substantiates these claims when naturalist guides make them? In a beautiful article from Science Magazine in 1976 (Science, New Series, Vol. 192, No. 4238 (Apr. 30, 1976), pp. 465-467), Kimberly Bailey tells us of her efforts to determine the ages of the islands of Santa Cruz, San Cristobal and Espanola. We will now use her data to find out the K-Ar ages of several of these islands ourselves.

(d) Read Bailey's short paper and give a brief synopsis (1 paragraph) of her approach and findings.

(e) Use the results from Sample H70-130 and JD1088 of Table 1 from Bailey's paper to determine ages for Santa Cruz Island and Santa Fe Island. To do this, you will need to navigate a few subtleties. First, note that the amount of Argon is presented in moles, and so you can use those numbers directly. To determine the number of moles of ^{40}K, you will need to use the weight percentage that is K_2O and use that in combination with the mass of the sample to figure out how much K is present. Note that not all of the potassium in the sample will be the isotope ^{40}K, so you will need to use the ratio of ^{40}K to total potassium, $\frac{^{40}\mathrm{K}}{\mathrm{K_{total}}} \approx 1.2 \times 10^{-4}$. Additionally, use the decay constant $\lambda \approx 5.8 \times 10^{-11}$ yr^{-1}.

Determining Lucy's age

In 1974, a fossil of *Australopithecus afarensis* (shown in Figure 1) was discovered in Ethiopia. This specimen, which was dubbed "Lucy," marks an important step in understanding human evolution because at the time of its discovery, it was the earliest known species to show evidence of bipedal locomotion. Because Lucy was found in an area that was rich in volcanic rock, potassium-argon dating was an ideal method for determining Lucy's age (Aronsen 1977).

Unfortunately for us, real-world K-Ar dating data are generally not neatly presented in the form of N_{Ar} and N_{K}. Instead, geologists will measure a concentration of ^{40}Ar in mol/g and a weight percent of K_2O. These data must be used to identify the number of ^{40}Ar and ^{40}K

nuclei in the sample. In this part of the problem, we will look at such measurements from an actual paleontological specimen as reported in Aronsen (1977) in order to determine its age.

Figure 1: The remains of Lucy, a specimen of *Australopithecus afarensis*.

(f) Using the table of ^{40}Ar and K_2O measurements below (Aronsen 1977), obtain an estimate for Lucy's age. Be sure to explain the steps you take to obtain your answer. Since each sample is taken from the area in which Lucy was found, we expect each sample to give you roughly the same answer; you will need to take the mean of the ages of each sample to obtain an estimate for Lucy's age.

Assume that each sample has a total mass of 1 g. Also, note that not all of the potassium in the sample will be the isotope ^{40}K, so you will need to use the ratio of ^{40}K to total potassium, $\frac{^{40}\mathrm{K}}{\mathrm{K_{total}}} \approx 1.2 \times 10^{-4}$. Additionally, use the decay constant $\lambda \approx 5.8 \times 10^{-11}$ yr^{-1}.

Sample Number	**^{40}Ar $\times 10^{-12}$ mol/g**	**wt.%K_2O**
1	2.91	0.657
2	3.18	0.755
3	3.08	0.680

Table 1: Outcome of measurements of potassium and argon for dating the rocks in the vicinity of Lucy.

2 Taylor series

In class, we solved the master equation for mRNA production and concluded that the mRNA production in steady state can be described by a Poisson distribution. To make this possible, we had to invoke the result that

$$\sum_{m=0}^{+\infty} \frac{1}{m!} x^m = e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \ldots \tag{3}$$

In this problem, we introduce the Taylor expansion to prove that the equation above is correct. This expansion is perhaps one of the most important tools used in the mathematical analysis of physical models.

(a) Read the section "The Math Behind the Models: The Beauty of the Taylor Expansion" on page 215 of PBoC2 shown below in Figure 2.

(b) The idea behind Equation 3 is that, as we sum more of the terms in the equation, our summation will converge to the function e^x. Here, we check this assertion using your favorite programming language. Make a plot like that shown in Figure 5.22 of PBoC (shown below in Figure 3), but for the function e^x. Specifically, plot the function e^x as well as the sum in the equation up until different powers. This means that you will plot e^x, together with 1, $1 + x$, $1 + x + \frac{1}{2}x^2$, etc. Go until the fourth order for a total of five lines on your plot.

3 Every Distribution Has Its Moments: The Poisson Distribution

In class, our null hypothesis for the mRNA distribution of a constitutive promoter led us to the Poisson distribution given by

$$p(m) = \frac{\lambda^m}{m!} e^{-\lambda}, \tag{4}$$

with $\lambda = r/\gamma$, r the rate of transcription and γ the rate of mRNA degradation. In this problem, we explore this distribution and its moments, namely, how the λ parameter dictates the mean and the variance of the Poisson distribution.

(a) Show that the Poisson distribution is normalized. To make this possible, you will have to invoke the Taylor expansion of the exponential function.

(b) Calculate the mean of the Poisson distribution, which is defined as

$$\langle m \rangle = \sum_{m=0}^{+\infty} m\, p(m). \tag{5}$$

To make this possible, you can use a trick that invokes the derivative. Specifically, note that

$$\lambda \frac{d}{d\lambda} \lambda^m = m\lambda^m \tag{6}$$

The Math Behind the Models: The Beauty of the Taylor Expansion A very important tool invoked in the mathematical analysis of physical models is the use of the so-called Taylor expansion. Series expansions of this kind will be one of our primary mathematical tools in the remainder of this book. The idea is very simple and amounts to replacing a function $f(x)$ in some neighborhood with a simple polynomial. As will be seen repeatedly throughout this book, the virtue of these approximations is that they allow us often to replace intractable nonlinear expressions with simple algebraic surrogates that we can handle analytically and that give an intuitive sense of the mathematics.

The idea of the Taylor expansion is embodied in the simple formula

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots . \tag{5.18}$$

Most of the time, we will only keep terms up to second order, and as a result the Taylor series algorithm reduces to the question: what three coefficients a_0, a_1, and a_2 should we use to best approximate the function $f(x)$?

For concreteness, let us consider the case in which we are interested in the behavior of the function $f(x)$ near $x = 0$. If we set $x = 0$ on both sides of Equation 5.18, we see that $a_0 = f(0)$. But we already know the function $f(x)$, so all we have to do is find its value at $x = 0$ to obtain the first coefficient. Next, let us take the derivative of both sides of Equation 5.18 with respect to x. We are left with the equation

$$f'(x) = a_1 + 2a_2 x + \cdots . \tag{5.19}$$

Once again, if we set $x = 0$, we are left with $a_1 = f'(0)$. We can continue to play the same game, this time evaluating the second derivative, with the result

$$f''(x) = 2a_2 + \cdots ; \tag{5.20}$$

which leads to $a_2 = \frac{1}{2} f''(0)$. This same basic analysis can be carried on indefinitely if one is interested in higher-order terms. Most of the time we will be content with the expression

$$f(x) \approx f(0) + f'(0)x + \tfrac{1}{2} f''(0)x^2 . \tag{5.21}$$

The symbol $\approx$ refers to the fact that in the neighborhood of the point x, the left- and right-hand sides of this equation are *approximately* equal. The conclusion of this little analysis is that if we want to find a simple quadratic surrogate for our function ofinterest, all we need to know is the value of the function and its first two derivatives at the point around which we are expanding. An example of this kind of analysis for the case of cos x is shown in Figure 5.22. In particular, using the rules given above, the Taylor series for this function is given by

$$\cos x \approx 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \frac{x^{10}}{10!} + \cdots . \tag{5.22}$$

Figure 5.22 compares the function cos x with various approximations based upon the Taylor series. We see that as more terms are included, the approximation is good for a wider range of values of x. Of course, there are mathematical subtleties that arise when considering a generic function, such as the question of convergence of the Taylor series. For example the function $1/(1-x)$ has the Taylor series, $1 + x + x^2 + x^3 + \cdots$, which is finite only for values of x such that $-1 < x < 1$.

Figure 2: Math Behind the Models: The Beauty of the Taylor Expansion. From PBoC2, page 215.

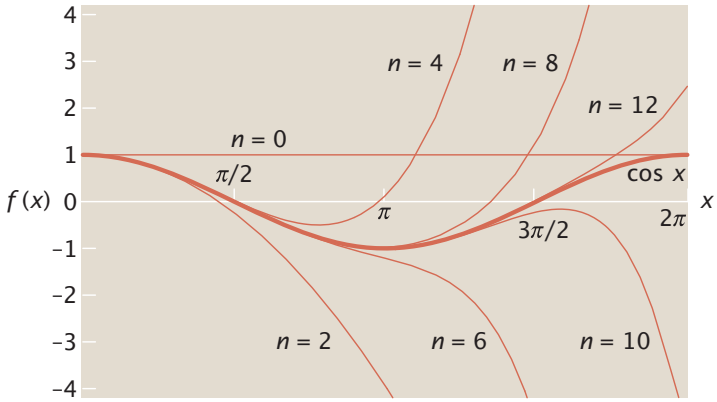

Figure 5.22: Comparison of the function $\cos x$ and its Taylor expansion. The curves are labeled by the order of the highest term kept in the Taylor series. For example, $n = 2$ means that the series goes to quadratic order, etc. The cosine function we are approximating is shown in bold for comparison with the approximate expressions.

Figure 3: Figure 5.22 from PBoC2.

in order to rewrite the sum for the mean as the derivate of a sum you know how to calculate. Is the result you get consistent with what you concluded when calculating the temporal evolution of the mean mRNA number in class?

Now, let's calculate the variance which is defined as

$$\text{var} = \text{SD}^2 = \langle (m - \langle m \rangle)^2 \rangle, \tag{7}$$

where SD is the standard deviation. The variance measures the width of the distribution by calculating its squared spread with respect to the mean.

(c) Show that the variance can be written as

$$\text{var} = \langle m^2 \rangle - \langle m \rangle^2. \tag{8}$$

To make this possible, you need to remember that the average of a number is just that number such that

$$\langle \langle m \rangle \rangle = \langle m \rangle, \tag{9}$$

and that the mean of a sum is the sum of the means.

(d) Invoke the derivative trick you used to calculate $\langle m \rangle$ in order to compute $\langle m^2 \rangle$.

(e) Calculate the variance. How does it compare to the mean? This is the key feature of the Poisson distribution! Visualize this by plotting the Poisson distribution for three different mean values, noting how the width of the distribution changes as its mean is modulated.

4 Sequencing Depth and the Poisson Distribution

The main challenge in assembling the sequence of a genome is to avoid any sequence gaps. To make this possible, it is important to sequence the genome at a *depth*—also called coverage—that ensures a minimal number of these gaps. In this problem, we explore how we can use the

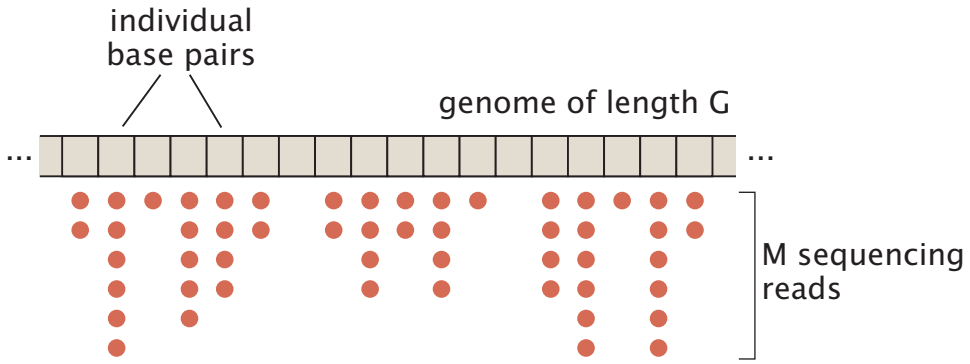

Figure 4: Simple model of genome sequencing. The base pairs of a genome of length G can be sequenced by any of the M sequence reads.

Poisson distribution to calculate this required sequencing depth in the context of a simple model of DNA sequencing. Specifically, Figure 4 shows how the G base pairs in a genome can be read out by M sequencing reads. As shown in the figure, some base pairs will be read out by the sequencing reads multiple times, while some base pairs will not be read out at all as predicted by the Poisson distribution.

(a) Given a genome of length G, and a total number of sequenced bases M, what is the mean number of reads per base pair λ?

(b) Let's assume that λ is small such that the number of reads falling on a given base pair is dictated by the Poisson distribution. What is the likelihood of that given base pair not being read at all?

(c) If you ask that the likelihood of a base pair not being read is 1%, what should the value of λ be? What does that say about the coverage, namely how many times the genome needs to be sequenced over in order to ensure this value of λ?

(d) Look up an example of sequencing coverage in your favorite DNA sequencing application. Do the numbers make sense given your calculation?

5 Mutation Per Generation in Humans

Comparing genetic sequences has served as a useful tool for determining how various organisms are related to each other. With the advent of the "genomic era," we no longer have to infer how living organisms are related to each other based on morphological traits alone. In this problem, we will begin to get a sense of the time scales over which mutations accumulate in genetic sequences and how we can use these mutations as a molecular clocks for determining the relationships between various organisms.

In this problem, we are ultimately interested in estimating the total number of mutations that are passed on in each human generation. As a first step, we must estimate the number of mutations that accumulate in a single cell division.

(a) Given that the human genome is 3 billion basepairs long and is replicated with an incredible fidelity of only one error in every 10^{10} basepairs per replication, how many mutations do you expect to see after one genome duplication?

With this number of mutations per genome duplication in hand, we can next tackle how many mutations are passed on by a mother and a father. Recall that while many mutations may occur in a given human, only those that accumulate in the gametes (egg and sperm) will actually be passed on. To determine the number of mutations that we expect to be passed on, we will need to consider the formation of the egg and the sperm separately as males and females have different developmental pathways regarding gametogenesis (see Figure 2).

As a primer for thinking about gametogenesis, let's briefly review the difference between mitosis and meiosis. Mitosis is the process by which a somatic cell duplicates its genome and then divides into two cells. Thus in a human, mitosis yields two cells with 46 chromosomes each. Meiosis, however, is the process by which a cell duplicates its genome and then proceeds to undergo two cell divisions, ultimately resulting in four cells with 23 chromosomes. This means that each round of mitosis requires one genome duplication and each round of meiosis requires one genome duplication (despite having two cell divisions).

In humans, females are born with all of their eggs nearly fully developed and they produce no new egg cells throughout the rest of their life. As illustrated in the top half of Figure 2, every developed egg is the result of 22 rounds of mitosis and 1 round of meiosis, yielding a total of 23 genome replications. This means that every egg a woman produces has undergone 23 genome replications regardless of a woman's age.

(b) Given the 23 genome duplications that occur in the process of forming an egg, how many mutations do you expect a woman to pass on to her children?

By contrast, spermatogenesis occurs continually throughout a male's lifetime upon reaching sexual maturity (i.e. puberty). At a bare minimum, a developed sperm cell has undergone 34 rounds of mitosis (30 leading to the formation of the stem cell and 4 after the stem cell) and 1 round of meiosis. But there are also additional rounds of mitosis to take into account as the result of the stem cells continually dividing to maintain the sperm supply. With these stem cells dividing every 16 days after puberty, the number of genome duplications to make a man's sperm is dependent on the age of the man.

(c) How many genome replications have occurred to make a "typical" man's sperm? In this context, we consider that a "typical" male hits puberty at 15 and reproduces at 30 years old.

(d) Given your answer in (c), how many mutations do you expect this "typical" man to pass

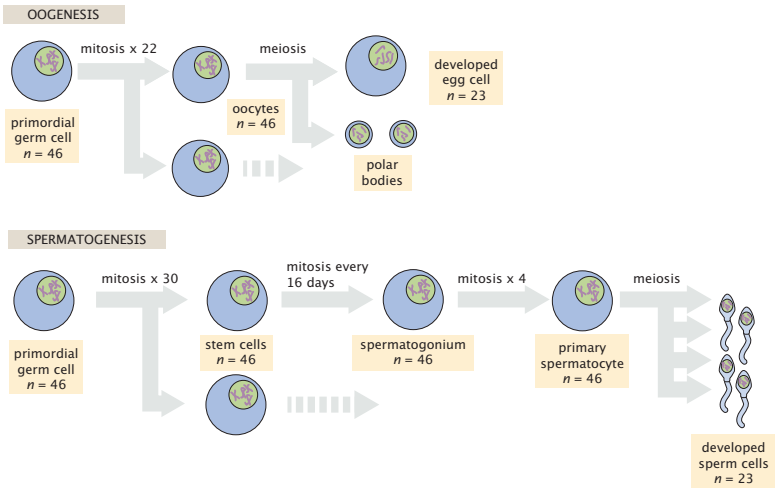

Figure 5: Schematic of oogenesis and spermatogenesis in humans. n refers to the number of chromosomes, where somatic cells have 46 and gametes have 23. For simplicity, the dashed arrows indicate the lineages of cells that we do not follow.

on to his children?

We have now estimated the total number of mutations that we expect the mother and the father to contribute, allowing us to determine the total number of mutations per human offspring.

(e) What is the total number of mutations we expect to accumulate in a human offspring? What are the relative effects of the mother and the father in this estimate?

(f) Make a plot of the number of mutations accumulated in the gametes as function of age for males and females. Make sure to graph the number of mutations in the egg and the sperm on the same plot to better compare their relative effects.