PHYS3511-Biological PhysicsFall 2018, Assignment #2

Read Chapter 3 of textbook before attempting the assignment

Exercise 1) The genome of the HIV–1 virus, like any genome, is a string of "letters" (base pairs) in an "alphabet" containing only four letters. The message for HIV is rather short, just $n = 10^4$ letters in all. Since any of the letters can mutate to any of the three other choices, there's a total of 30000 possible distinct one-letter mutations.

In 1995, A. Perelson and D. Ho estimated that every day about 10^{10} new virus particles are formed in an asymptomatic HIV patient. They further estimated that about 1% of these virus particles proceed to infect new white blood cells. It was already known that the error rate in duplicating the HIV genome was about one error for every 3×10^4 "letters" copied. Thus, the number of newly infected white cells receiving a copy of the viral genome with one mutation is roughly

$$10^{10} \times 0.01 \times (10^4/3 \times 10^4) \approx 3 \times 10^7$$

per day. This number is much larger than the total 30000 possible 1-letter mutations, so every possible mutation will be generated several times per day.

A) How many distinct two-base mutations are there? **Hint:** see class notes.

In class we defined the number of base(bp)/nucleotide(nt) as $n = 10^4$, and the number of mutations as i, and the **total number** of i-base mutations as:

$$N_i^{total} = \frac{3^i n(n-1)\ldots(n-i+1)}{i!} = \frac{3^i n!}{i!\,(n-i)!} = 3^i \binom{n}{i}$$

Hence the total number of 2-base mutation is:

$$N_2^{total} = \frac{3^i 10000(10000-1)}{2!} = 4 \times 10^8$$

B) You can work out the probability P_2 that a given viral particle has two bases copied inaccurately from the previous generation using the sum and product rules of probability. Let $p = \frac{1}{3\times10^4}$ be the probability that any given base is copied incorrectly. Then the probability of exactly two errors is P_2, times the probability that the remaining 9998 letters don't get copied inaccurately, times the number of distinct ways to choose which two letters get copied inaccurately. Find P_2.

The **probability** of a i-base mutation each time an HIV is replicated is:

$$P_i = \binom{n}{i} p^i (1-p)^{n-i},$$

where $p = \frac{1}{3\times10^4}$ is the probability of a RNA copy error each time an HIV is replicated. For a 2-base mutation:

$$P_2 = \binom{10000}{2} p^2 (1-p)^{10000-2} = 0.04,$$

C) Find the expected number of two-letter mutant viruses infecting new white cells per day and compare to your answer to (A). Hence estimate how long it would take HIV to mutate to a form that is resistant to **two anti-virial drugs** that deactivate **two amino acids sites of HIV**. **Hint:** see class notes.

Following the logic used for 1-base mutation the number of i-base mutation per day is:

$$N_i = 10^{10} \times 0.01 \times P_i.$$

So the number of 2-base mutations per day is

$$N_2 = 10^{10} \times 0.01 \times P_2 = 4 \times 10^6 day^{-1}.$$

The time it would take for HIV to mutate to a 2-base drug resistant form is simply:

$$time = \frac{N_2^{total}}{N_2} \sim 112 days.$$

D) Repeat (A–C) for **three independent mutations**.
the total number of 3-base mutation is:

$$N_3^{total} = \frac{3^i 10000(10000-1)(10000-2)}{3!} = 4.5 \times 10^{12}$$

For a 3-base mutation:
$P_3 = \binom{10000}{3} p^3 (1-p)^{10000-3} = 0.0044$, and
$N_3 = 10^{10} \times 0.01 \times P_3 = 4.4 \times 10^5 day^{-1}$.
The time it would take for HIV to mutate to a 2-base drug resistant form is simply:

$$time = \frac{N_3^{total}}{N_3} \sim 1 \times 10^7 days$$

E) Suppose an antiviral drug attacks some part of HIV, but that the virus can evade the drug's effects by making one particular, single-base mutation. According to the information above, the virus will very quickly stumble upon the right mutation—the drug isn't effective for very long. Why do you suppose current HIV therapy involves a combination of **three** different antiviral drugs **simultaneously administered**?
For a combination of two antivirial drugs that deactivate two sites, the result of part c shows that after about 112 days (half a year) HIV would have mutate to a form resistant to the drugs. In contrast the result of part D shows that it would take centuries for HIV to mutate to a from resistant to three drugs simultaneously.

Exercise 2) <u>Problem 3.1 of Chapter 3</u>

3.1 Growth and the logistic equation

In this chapter we described the logistic equation as a simple toy model for constrained growth of populations. The goal is to work out the dynamics in more detail.

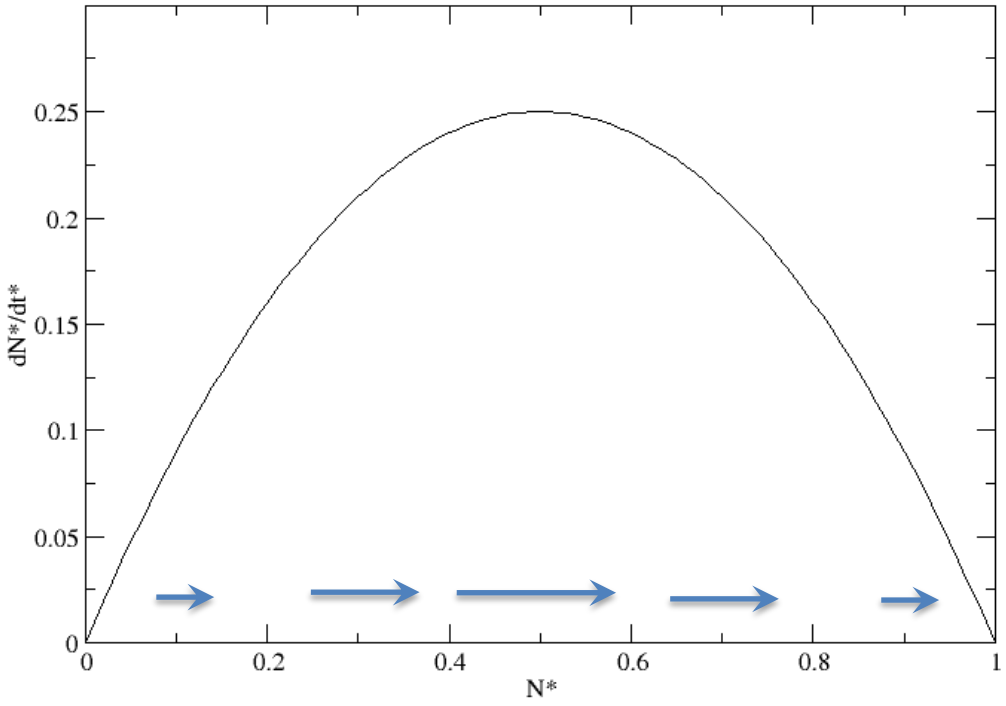
- **A)** Rewrite the equation in dimensionless form and explain what units this means time is measured in.
  Equation 3.8 is $\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right)$, where r is the **rate** in unit of $particle \cdot s^{-1} = s^{-1}$, and of course, $N(t)$ is the population or the number of particles (i.e. proteins, mRNA, etc.), and K is the "maximum population". We can define the "growth time" $\tau = 1/r$ in unit of s. Let's rewrite the equation by dividing both sides by rK:
  $$\frac{d(N/K)}{d(rt)} = \frac{N}{K}\left(1 - \frac{N}{K}\right) \rightarrow \frac{dN^*}{dt^*} = N^*(1 - N^*),$$
  where $N^* = N/K$is the scaled population, so that N* is the population measured in unit of the maximum population, K. and $t^* = rt = t/\tau$ is the dimensionless time. t* is time in measured in units of the "growth time" $\tau = 1/r$.
- **B)** Find the number of N at which the **net growth rate** is maximum.
  I believe that there is a misprint in the question, and the authors really meant **net growth rate** (see figure 3.10).

The **net growth rate** is $\frac{dN^*}{dt^*}$. To **maximize** we use **basic calculus**, and **differentiate** the net growth rate with respect to N*, and set it to **zero**. To do so write the net growth rate as a function of the scaled population, $\frac{dN^*}{dt^*} = f(N^*) = N^*(1 - N^*)$. Find the maximum by

setting differentiation with respect to $\frac{df(N^*)}{dN^*} = 0 \rightarrow 1 - 2N^* = 0 \rightarrow N^*_{max}$=0.5. To prove that it is a maximum you must find the second derivative and evaluate it at $N^* = N^*_{max}$=0.5, we find the second derivative $\frac{d^2f(N^*)}{dN^{*2}} = -2 < 0$. Hence using the rules of calculus it is a maximum.

C) To find the maximum value just evaluate the net growth rate at 0.5.
$\frac{dN^*}{dt^*}(N^*_{max}) = \frac{dN^*}{dt^*}(0.5) = 0.5(1 - 0.5) = 0.25.$

D) To do this one show that $\frac{dN^*}{dt^*} = N^*(1 - N^*)$ can be rewritten as that $\frac{dN^*}{dt^*} = -a(N^* - N^*_{max})^2 + b$, where you must find a, and b. This is just an inverted parabola.
$\frac{dN^*}{dt^*} = -(N^* - 0.5)^2 + 0.25$. This is an inverted parabola with a maximum at $N^* = N^*_{max}$=0.5, and $\frac{dN^*}{dt^*} = 0$, at $N^* = 0\ and\ 1$. You should be able to do a plot similar to Fig 3.10 now:



Exercise 3) <u>Problem 3.2 of Chapter 3</u>

3.2 Improve the protein synthesis rate of E. Coli during a cell cycle from section 3.1.3 and 3.2.1 by including the effect of protein degradation. For simplicity, assume that all proteins are degraded at the same rate with a half-life of 60 minutes and work out and work out the number of ribosomes needed to produce the protein content of a new bacterium given that part of the synthesis is required for the replacement of degraded proteins. Compare with the results from the estimates in section 2.1.1 (page 38)
On page 109 (section 3.2.1), the authors estimate the protein production rate as follows. There are $N_{ribosomes} = 2 \times 10^4$ ribosomes in an E. Coli, and each ribosome can produce proteins at a rate of $r_{growth} = \frac{1}{20} s^{-1}$(here the authors state that a ribosome can produce a

protein in about 20 seconds. This gives a **net growth rate** of $\beta_{growth} = N_{ribosomes} r_{growth} = 1000 \frac{proteins}{s}$. If we write the protein rate equation as:

$$\frac{dN_p}{dt} = \beta_{growth},$$

Then the number of proteins, N_p, after the cell division of $\tau_{division} \approx 3000s$ is simply:

$$\frac{dN_p}{dt} = \beta_{growth} \rightarrow N_p = \tau_{division}\beta_{growth} + N_{p0} = 3 \times 10^6,$$

where the authors assumed that the initial number of proteins (destined for the second cell) is zero, $N_{p0} = 0$.

If we now assume that proteins decay with a half-life of 60 minutes or 3600s, then the physics of the problem is altered. As mentioned in the textbook, and in class, proteins are degraded by proteases, which are proteins. The rate of decay must be proportional to the number of the protein present, $\frac{dN_p}{dt} = -k_{decay} N_p$, with solution:

$$\frac{dN_p}{dt} = -k_{decay} N_p \rightarrow N_p = N_{p0} exp\left(-k_{decay} t\right),$$

where N_{p0} is the initial number of protein at t =0. The half life, $t_{1/2}$, is the time that it takes for the number of proteins to be reduced to half its initial amount:

$$N = 0.5N_{p0} \rightarrow N_p = 0.5N_{p0} = N_{p0} exp\left(-k_{decay} t_{1/2}\right) = t_{1/2} = \frac{\ln 2}{k_{decay}} \rightarrow k_{decay} = \frac{ln\, 2}{t_{1/2}},$$

with $t_{1/2}$ = 3600s. All this assumes that the only process is protein decay.

However, if we were to combined ribosomes production of proteins with protein decay, the **differential equation (DE)** become:

$$\frac{dN_p}{dt} = \beta_{growth} - k_{decay} N_p,$$

We will assume a solution of the form

$$N_p = N_{max}\left(1 - exp\left(-k_{decay} t\right)\right)$$

Verify by direct substitution to left hand side (LHS)

$$\frac{d\left(N_{max}\left(1 - exp\left(-k_{decay} t\right)\right)\right)}{dt} = N_{max} k_{decay} exp\left(-k_{decay} t\right) = \beta_{growth} - k_{decay} N_p.$$

Then to right hand side

$$N_{max} k_{decay} exp\left(-k_{decay} t\right) = \beta_{growth} - k_{decay} N_{max}\left(1 - exp\left(-k_{decay} t\right)\right),$$

and with some rearrangement

$$N_{max} k_{decay} exp\left(-k_{decay} t\right) = \beta_{growth} - k_{decay} N_{max} + k_{decay} N_{max} exp\left(-k_{decay} t\right),$$

which gives

$$0 = \beta_{growth} - k_{decay} N_{max} \rightarrow N_{max} = \frac{\beta_{growth}}{k_{decay}}$$

with the maximum number of proteins being: $N_{max} = \frac{\beta_{growth}}{k_{decay}}$.

Unit analysis (get units from above) gives $N_{max} = \frac{\beta_{growth}}{k_{decay}} = \frac{proteins \cdot s^{-1}}{s^{-1}}$ = proteins.

We can use this to evaluate $k_{decay} = \frac{ln\,2}{t_{1/2}} = \frac{ln\,2}{3600s} = 1.925 \times 10^{-4} s^{-1}$, and if we assumed that the number of proteins must still be $N_{max} = 3 \times 10^6$, and $\beta_{growth} = N_{ribosomes} r_{growth}$, with $r_{growth} = \frac{1}{20} s^{-1}$.

$$N_{max} = \frac{\beta_{growth}}{k_{decay}} \rightarrow 3 \times 10^6 = \frac{N_{ribosomes} \times \frac{1}{20} s^{-1}}{1.925 \times 10^{-4} s^{-1}} \rightarrow N_{ribosomes} = 1.155 \times 10^4$$

If students do all of the above they can get full grades.

Bonus: Some students may note that $N_{ribosomes} = 1.155 \times 10^4$ is actually less than the book's original estimate of 2×10^4. This does not make sense since if protein degradation is taken into account, then more proteins than the final value of 3×10^6 must be produced, which means that the number of ribosomes must be higher than 2×10^4. To obtain bonus mark students must do the following. If after t = 3000s the number of proteins must be $N_p = 3 \times 10^6$, then we must have:

$$N_p = N_{max}\left(1 - exp\left(-k_{decay} t\right)\right) \rightarrow 3 \times 10^6 = N_{max}\left(1 - exp\left(-k_{decay} t\right)\right),$$

where $k_{decay} = \frac{ln\,2}{t_{1/2}} = \frac{ln\,2}{3600s} = 1.925 \times 10^{-4} s^{-1}$, and t = 3000 s. Solving for N_{max}:

$$N_{max} = 6.8 \times 10^6.$$

Using $N_{max} = \frac{\beta_{growth}}{k_{decay}} \rightarrow 6.8 \times 10^6 = \frac{N_{ribosomes} \times \frac{1}{20} s^{-1}}{1.925 \times 10^{-4} s^{-1}} \rightarrow N_{ribosomes} = 2.6 \times 10^4$,

which I will note is greater than the 20000 estimate in the book.

Exercise 4) <u>Problem 3.3 of Chapter 3</u>

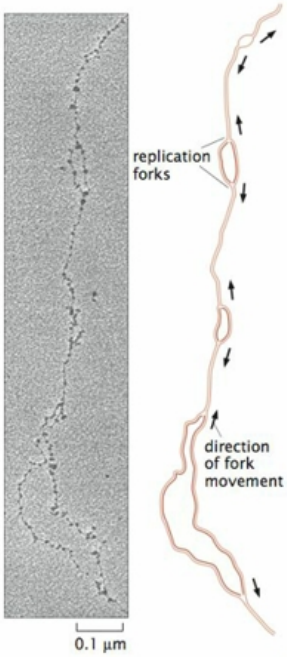


Figure 3.35: Replication forks in *D. melanogaster*. Replication forks move away in both directions from replication origins. (Electron micrograph courtesy of Victoria Foe. Adapted from B. Alberts et al., Molecular Biology of the Cell, 5th ed. Garland Science, 2008.)

• 3.3 DNA replication rates

Assuming that Figure 3.35 is a representative sample of the replication process:

(a) Estimate the fraction of the total fly genome shown in the micrograph. Note that the fly genome is about 1.8×10^8 nucleotide pairs in size.

(b) Estimate the number of DNA polymerase molecules in a eukaryotic cell like this one from the fly *D. melanogaster*.

(c) There are eight forks in the micrograph. Estimate the lengths of the DNA strands between replication forks 4 and 5, counting up from the bottom of the figure. If a replication fork moves at a speed of roughly 40 bp/s, how long will it take for forks 4 and 5 to collide?

(d) Given the mean spacing of the bubbles, estimate how long it will take to replicate the entire fly genome.

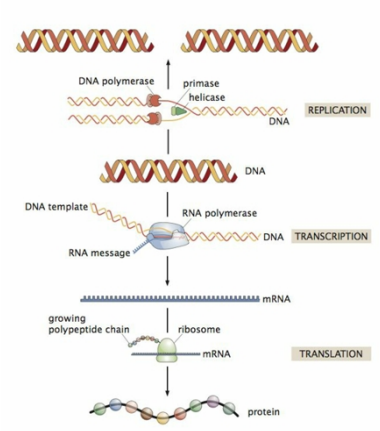
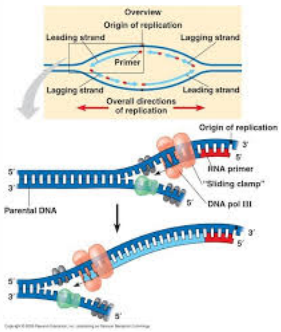


Figure 3.12: The processes of the central dogma. DNA is replicated to make a second copy of the genome. Transcription refers to the process in which RNA polymerase makes an mRNA molecule. Translation refers to the synthesis of a polypeptide chain whose sequence is dictated by the arrangement of nucleotides on mRNA.



3.3 a) A rough estimate of fig. 3.35 shows that the length of the visible DNA portion is $2\mu m$. Using data from table 1.1 (see last page of this document) the length of 1 basepair (bp) is $\sim\frac{1}{3}nm = \frac{1}{3}\times 10^{-3}\mu m$. This gives the number bp of the DNA portion is $\frac{2\mu m}{\frac{1}{3}\times 10^{-3}\mu m\cdot bp^{-1}} = 6000bp$. The **fraction of the fly genome** is $\frac{6000bp}{1.8\times 10^{8}bp} = \frac{1}{3}\times 10^{-4}$.

b) The above shows figure 3.12 from the textbook of a DNA replication fork with **two DNA polymerase**, which I note is actually a big simplification of the actual process. To estimate the number of DNA polymerase in the cell, assume that all polymerases are attached to the DNA during replication. Since there are 4 bubbles in the DNA portion of 3.35, there are 8 polymerases in that DNA portion. This gives a total number of polymerases $\frac{8}{\frac{1}{3}\times 10^{-4}} = 240000$.

c) The distance between for 4 and 5 is about $0.5\mu m$, or converting $\frac{0.5\mu m}{\frac{1}{3}\times 10^{-3}\mu m\cdot bp^{-1}} = 1500bp$. If a replication fork moves at $40\frac{bp}{s}$, then two adjacent forks should move together at a rate of $80\frac{bp}{s}$, then fork 4 and 5 should collide after, $\frac{1500bp}{80\frac{bp}{s},}$=18.75s.

d) Counting from the bottom up the distance from the center of bubble 1 to bubble 2 is $0.8\mu m$, and bubble 2 and 3 is $0.7\mu m$, and bubble 3 and 4 is $0.6\mu m$. Hence, the average distance is $0.7\mu m$, or $\frac{0.7\mu m}{\frac{1}{3}\times 10^{-3}\mu m\cdot bp^{-1}} = 2100bp$.. The estimate of the DNA replication time is$\frac{2100bp}{80\frac{bp}{s},}$=26.25s

Table 1.1: Rules of thumb for biological estimates.

	Quantity of interest	Symbol	Rule of thumb
E. coli			
	Cell volume	$V_{E.\,coli}$	$\approx 1\,\mu m^3$
	Cell mass	$m_{E.\,coli}$	≈ 1 pg
	Cell cycle time	$t_{E.\,coli}$	≈ 3000 s
	Cell surface area	$A_{E.\,coli}$	$\approx 6\,\mu m^2$
	Macromolecule concentration in cytoplasm	$c_{E.\,coli}^{\text{macromol}}$	≈ 300 mg/mL
	Genome length	$N_{\text{bp}}^{E.\,coli}$	$\approx 5 \times 10^6$ bp
	Swimming speed	$v_{E.\,coli}$	$\approx 20\,\mu m/s$
Yeast			
	Volume of cell	V_{yeast}	$\approx 60\,\mu m^3$
	Mass of cell	m_{yeast}	≈ 60 pg
	Diameter of cell	d_{yeast}	$\approx 5\,\mu m$
	Cell cycle time	t_{yeast}	≈ 200 min
	Genome length	$N_{\text{bp}}^{\text{yeast}}$	$\approx 10^7$ bp
Organelles			
	Diameter of nucleus	d_{nucleus}	$\approx 5\,\mu m$
	Length of mitochondrion	l_{mito}	$\approx 2\,\mu m$
	Diameter of transport vesicles	d_{vesicle}	≈ 50 nm
Water			
	Volume of molecule	V_{H_2O}	$\approx 10^{-2}\,nm^3$
	Density of water	ρ	$1\,g/cm^3$
	Viscosity of water	η	≈ 1 centipoise (10^{-2} g/(cm s))
	Hydrophobic embedding energy	$\approx E_{\text{hydr}}$	2500 cal/(mol nm^2)
DNA			
	Length per base pair	l_{bp}	$\approx 1/3$ nm
	Volume per base pair	V_{bp}	$\approx 1\,nm^3$
	Charge density	λ_{DNA}	$2\,e/0.34$ nm
	Persistence length	ξ_p	50 nm
Amino acids and proteins			
	Radius of "average" protein	r_{protein}	≈ 2 nm
	Volume of "average" protein	V_{protein}	$\approx 25\,nm^3$
	Mass of "average" amino acid	M_{aa}	≈ 100 Da
	Mass of "average" protein	M_{protein}	$\approx 30{,}000$ Da
	Protein concentration in cytoplasm	c_{protein}	≈ 150 mg/mL
	Characteristic force of protein motor	F_{motor}	≈ 5 pN
	Characteristic speed of protein motor	v_{motor}	≈ 200 nm/s
	Diffusion constant of "average" protein in cytoplasm	D_{protein}	$\approx 10\,\mu m^2/s$
Lipid bilayers			
	Thickness of lipid bilayer	d	≈ 5 nm
	Area per molecule	A_{lipid}	$\approx \frac{1}{2}\,nm^2$
	Mass of lipid molecule	m_{lipid}	≈ 800 Da