MCB137L/237L: Physical Biology of the Cell
Spring 2020
Homework 10:
(Due 4/23/20 at 3:30pm)

Hernan G. Garcia

1 Physical Biology of Viruses

In this problem, we take a random walk through the physical biology of viruses, honoring them as one of the most sophisticated, interesting and scary parts of the biological world.

(a) Let's begin by considering the data storage capacity of viruses. Choose an RNA virus (such as influenza, HIV or COVID-19) and a bacteriophage (such as lambda or T4) and compute the physical data storage capacity when their genomes are packed within the virion. Figure 1 is a resource that will allow you to understand viral sizes. I am talking about units of bits/μm^3. Compare this to the 4 TB hard drive that one uses to back up a laptop. How many viruses would it take to store the entirety of the Library of Congress? How much volume would all of those viruses take up?

To give you some ideas for how to tackle this problem, assume that each printed character is 8 bits (1 byte) while a base pair is 2 bits (remember that you can come up with four possible unique 2-bit words, corresponding to the four nucleotides, and that, because we are working with DNA in this problem, each nucleotide is complemented with another one, so we assume no additional information is encoded on the other strand of DNA).

(b) One of the most important properties of a given infection is the so-called "burst size", the number of new viruses produced per infected cell. One of the original hypotheses (which you will refute here) for what controls the burst size is the available volume within the host cell. Given that for a typical bacteriophage infection the burst size is roughly 100 viruses, what fraction of the volume is taken up by the newly synthesized viruses? Figure 2 shows an electron microscopy image of an infected bacterium.

(c) How are viruses transmitted? Three key routes are through the respiratory tract, the digestive tract and the reproductive tract. In all three cases, our bodies are set up with a number of different tricks to resist infection including mucus and ciliary transport in our respiratory and digestive tracts and harsh conditions in our digestive tract such as low pH. The current coronavirus epidemic is apparently passed through the respiratory tract and in

virus	size (nm)	genome size (base pairs)	genome type, capsid structure	BNID
porcine circovirus (PCV)	17	1,760	circular ssDNA, icosahedral	106467, 106468
cowpea mosaic virus (CPMV)	28	9,400	2 ssRNA molecules, icosahedral	106454, 106455
cowpea chlorotic mottle virus (CCMV)	28	7,900	3 ssRNA molecules, icosahedral	106456, 106457
φX174 (*E. coli* bacteriophage)	32	5,400	ssDNA, icosahedral	103246, 106442
tobacco mosaic virus (TMV)	40×300	6,400	ssRNA, rod shaped	104376, 104375, 106453
polio virus	30	7,500	ssRNA, icosahedral	103114, 111324
φ29 (*Bacillus* phage)	45x54	19,000	dsDNA, icosahedral (T3)	109734
lambda phage	58	49,000	dsDNA, icosahedral (with tail)	103122, 105770
T7 bacteriophage	58	40,000	dsDNA, 55 genes, icosahedral (T7)	109732, 109733
adenovirus (linear DNA)	88-110	36,000	dsDNA, icosahedral	103114, 103115, 106441
influenza A	80-120	14,000	ssRNA, roughly spherical	104073, 105768
HIV-1	120-150	9,700	ssRNA, roughly spherical	101849, 105769
herpes simplex virus 1	125	153,000	dsDNA, icosahedral	103114, 106458
Epstein-Barr virus (EBV)	140	170,000	dsDNA, icosahedral	103246, 111424
mimivirus	500	1,200,000	dsDNA, icosahedral	105142, 105143
pandora virus	500x1000	2,800,000	dsDNA, icosahedral	109554, 109556

Figure 1: Sizes of viruses. The table considers both RNA and DNA viruses and reports both the size of the virion and the length of the genome.

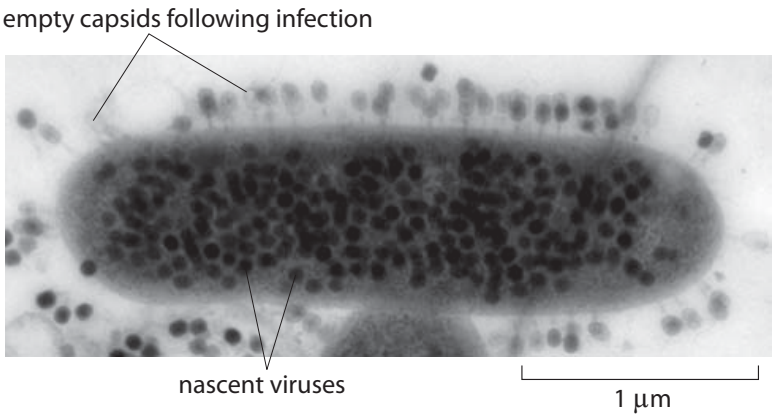

Figure 2: Synthesis of new viruses in an infected bacterium.

TABLE II. Numbers of Particles in Different Initial Diameter Ranges Emitted in One Cough and One Sneeze According to Duguid

Diameter Range (μm)	Number of Particles in a Cough	Number of Particles in a Sneeze
1–2	50	26,000
2–4	290	160,000
4–8	970	350,000
8–16	1600	280,000
16–24	870	97,000
24–32	420	37,000
32–40	240	17,000
40–50	110	9000
50–75	140	10,000
75–100	85	4500
100–125	48	2500
125–150	38	1800
150–200	35	2000
200–250	29	1400
250–500	34	2100
500–1000	12	1000
1000–2000	2	

Source: Data from Duguid, "The Size and Duration of Air-Carriage of Respiratory Droplets and Droplet-Nuclei." *Journal of Hygiene* 4:471–480, Table 3 (1946).

Figure 3: Distribution of droplet sizes after a sneeze.

this part of the problem, we appeal to Figure 3 for a look at the distribution of droplet sizes. How many particles are contained in a typical cough or sneeze? How much volume is that? Given a viral concentration in sputum of 10^6 to 10^{11} RNAs/ml, estimate how many virions of SARS-CoV-2 will be carried in a typical droplet. A very interesting source of information on this is the work of Prof. Lydia Bourouiba from MIT who does visualization experiments on humans coughing. https://www.nature.com/articles/d41586-019-00065-5 - an excellent brief interview with Bourouiba on the physics of sneezing and coughing.

2 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.

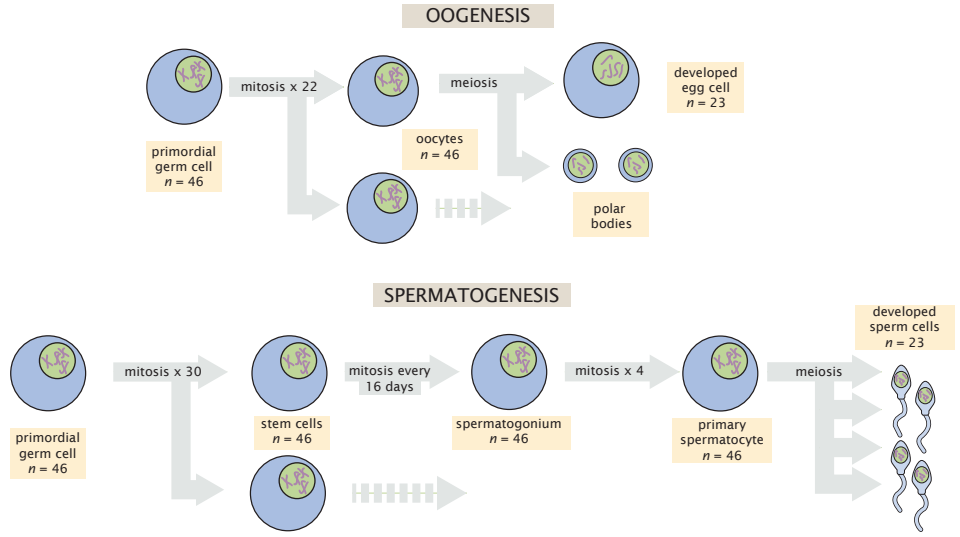

Figure 4: Schematic of oogenesis and spermatogenesis in humans. n refers to the number of chromosomes, where somatic cells have 46 and gametes have 23. For simplicity, the dashed arrows indicate the lineages of cells that we do not follow.

(d) Given your answer in **2c**, how many mutations do you expect this "typical" man to pass 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.

3 Open reading frames in random DNA

Do problem 4.7 of PBoC shown in Figure 5. Note that Figure 1.4 from PBoC can be found in Figure 6 of this homework. Finally, note that (b) and (c) in the problem are asking you to compute the probability of having an ORF of *at least* N codons in length.

4 Dynamics of the constitutive promoter

In class, we determined that the rate of mRNA decay γ, and not the production rate r, dictates the time it takes for the mean mRNA number to reach its steady state value. Here, we further explore this conclusion that could be at odds with our initial expectations about

• 4.7 Open reading frames in random DNA

In this problem, we will compute the probabilities of finding specific DNA sequences in a perfectly random genome, for which we assume that the four different nucleotides appear randomly and with equal probability.

(a) From the genetic code shown in Figure 1.4, compute the probability that a randomly chosen sequence of three nucleotides will correspond to a stop codon. Similarly, what is the probability of a randomly chosen sequence corresponding to a start codon?

(b) A reading frame refers to one of three possible ways that a sequence of DNA can be divided into consecutive triplets of nucleotides. An open reading frame (ORF) is a reading frame that contains a start codon and does not contain a stop codon for at least some minimal length of codons. Derive a formula for the probability of an ORF having a length of N codons (not including the stop codon).

(c) The genome of *E. coli* is approximately 5×10^6 bp long and is circular. Again assuming a that the genome is a random configuration of base pairs, how many ORFs of length 1000 bp (a typical protein size) would be expected by chance? Note that there are six possible reading frames.

(Problem courtesy of Sharad Ramanathan)

Figure 5: Problem 4.7 from PBoC.

the dynamics of the constitutive promoter.

(a) If r does not dictate the time to reach steady state, what aspect of the promoter dynamics does it determine? Solve for the mRNA concentration as a function of time for two different values of r such as 10 mRNA/min and 20 mRNA/min using an initial condition $m(t = 0) = 0$. Use $\gamma = 1$ /min. Plot mRNA number vs. time and show that r controls the initial slope.

(b) If r determines this initial slope, how come both curves take the same time to reach their steady state value? Plot the phase diagrams corresponding to both choices of model parameters and show that, while $r = 20$ mRNA/min has a faster initial increase in mRNA number, its steady state value is also larger such that the time it takes to reach steady state remains unaltered.

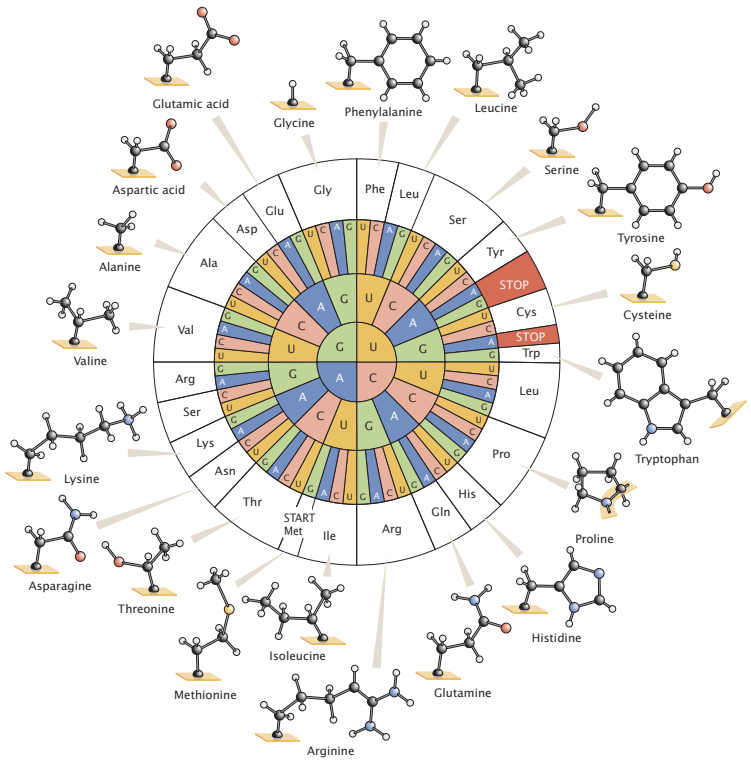

Figure 6: Genetic code. In this schematic representation, the first nucleotide in a coding triplet is shown at the center of the ring, the second nucleotide in the middle colored ring and the third nucleotide in the outer colored ring. In this representation of the genetic code, the four bases are adenine (A), cytosine (C), guanine (G) and uracil (U). Uracil is structurally very similar to thymine (T), and is used instead of thymine in messenger RNA. The amino acids corresponding to each group of triplets are illustrated with their names (outer ring) and atomic structures. Two amino acids, tryptophan and methionine, are encoded by only a single triplet, whereas others including serine, leucine, and arginine are encoded by up to six. Three codons do not code for any amino acid and are recognized as stop signals. The unique codon for methionine, AUG, is typically used to initiate protein synthesis.