BE/APh161: Physical Biology of the Cell
Homework 3
Due Date: Wednesday, January 31, 2024

"Champions aren't made in gyms. Champions are made from something they have deep inside them - a desire, a dream, a vision. They have to have the skill, and the will. But the will must be stronger than the skill." - Muhammad Ali

1. Energy and Life

One of the strongest things we can say about the properties of living organisms that distinguish them from inorganic materials such as the rocks that make up the face of Half Dome is that they are always consuming energy. Figure 1 shows a number of biological processes as viewed through the prism of energy consumption.

(A) Write a brief, thoughtful paragraph about the meaning of the energy scale k_BT.

(B) In this problem, choose three of the entries in the figure and make your own calculation of the relevant energy scale and see to what extent you agree with the reported numbers. Don't find a way to get the same numbers as are in the figure. Rather, do this yourself and get your own number. Make sure you carefully report your thought process and assumptions.

2. Phosphorus, Sulfur and the Lives of Cells

In addition to the big ticket chemical elements in cells (carbon, nitrogen, oxygen, hydrogen), other elements come in at lower concentrations, but still with enormous functional importance. Two such elements are phosphorus and sulfur and in this problem, we will try to figure out how much of the cell's dry weight is taken up by these elements and what this implies about the transport of these elements into the cellular interior.

(a) Let's begin by trying to estimate the number of phosphorus atoms in a cell. Where do we find phosphorus? There is 10 mM of ATP in a typical bacterium. We all know that in both RNA and DNA, every base carries its

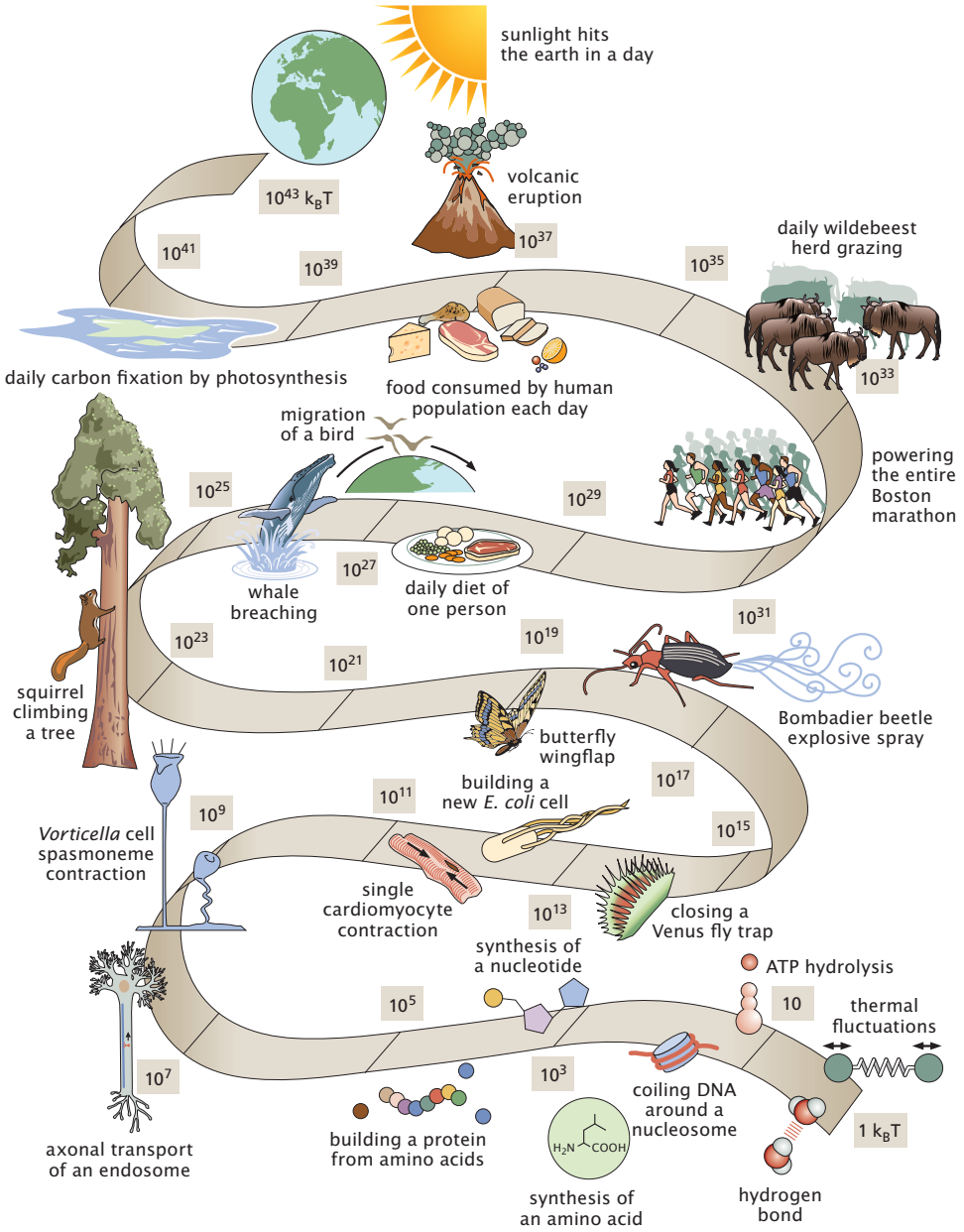

Figure 1: Energy scales of biology. From top to bottom, the energetic cost of the process of interest increases. All energies are measured in units of k_BT.

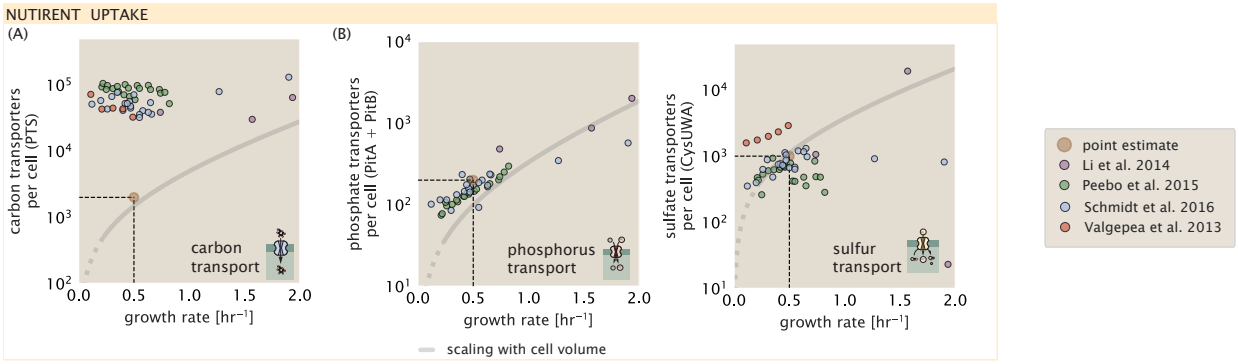

Figure 2: Census of carbon, phosphorus and sulfur transporters in *E. coli*.

own phosphate. Many lipids are phospholipids, with polar heads containing phosphate atoms as well. Proteins are phosphorylated. Don't forget ribosomes. They too are full of phosphorus atoms because they are 2/3 by mass RNA. Given these various facts, estimate the total number of phosphorus atoms in a bacterium. Given a division time of $f \times 10^3$ s, how many phosphate transporters (PitA) are needed to bring all those phosphorus atoms into the cell during that time?

(b) Next, we consider sulfur. Where do we find sulfur atoms in cells? Clearly one of the main amino acids, cysteine, has its known covalent binding properties precisely because of its sulfur atom. The metabolite glutathione has a concentration of 17 mM. Like in the previous part of the problem, in light of these facts, make an estimate of the total number of sulfur atoms in a bacterial cell. Given a division time of $f \times 10^3$ s, how many sulfur transporters (CysUWA) are needed to bring all those sulfur atoms into the cell during that time?

(c) How do your results from the first two parts compare to the measured numbers as reported in Figure 2.

3. Synthesizing a Transcriptome: Big Data in Transcription

In class, we briefly discussed the myriad of different ways to measure gene expression. Writ large, we can either find ways to count the mRNA

transcripts or the protein products that result from these transcripts. For example, when properly calibrated, the green fluorescent protein (GFP) in conjunction with fluorescence microscopy is a favorite approach for measuring protein copy numbers. Recently, a different way to engage in the dialogue between theory and experiment has been afforded by the advent of technologies that make it possible to take a census of the full complement of transcripts inside individual cells.

One of the key applications of single-cell mRNA sequencing has been its use to identify "transcriptional fingerprints" that define discrete cell types within a population containing cells that have committed to multiple possible fates. One of the best examples of this application of single-cell transcriptome-wide sequencing comes from projects such as the *Tabula muris*. This project measured RNA counts for tens of thousands of genes within tens of thousands of individual cells in the mouse, derived from tens of distinct organs and tissues. Each single cell transcriptome is a giant $\approx$ 10,000 dimensional vector with the i^{th} entry corresponding to the mRNA count of the i^{th} gene.

One widespread approach to visualizing the results from these types of experiments is shown in Figure 3. In the figure, each point corresponds to an individual cell whose transcriptome was sequenced. Here, the extremely high dimensional data resulting from single-cell RNA sequencing (i.e., the number of mRNA molecules corresponding to each of $\approx$ 10,000 genes in each cell) was projected onto two dimensions using methods we will later explore. Further, once this projection is performed, cells are grouped in clusters. The idea is that cells within a cluster share much of their gene expression profile and are therefore identified as unique cell types corresponding to different tissues within the mouse. In this problem, we will attempt to build some intuition for how this identification of unique cell types is achieved by working with a synthetic transcriptome that we build ourselves using our understanding of the constitutive promoter. Obviously this is a caricature of the real situation where most genes are *not* constitutively expressed.

(A) Let's start by creating a mental picture of the high dimensionality of single-cell sequencing data by picturing how this data is stored. Specifically, think of a matrix $\mathbf{G}$ where you store the RNA counts for 10,000 genes measured in 1,000 cells where each row of the matrix corresponds to a given cell. How many rows and columns would this matrix have? Draw this matrix schematically, clearly indicating what each dimension of the matrix repre-

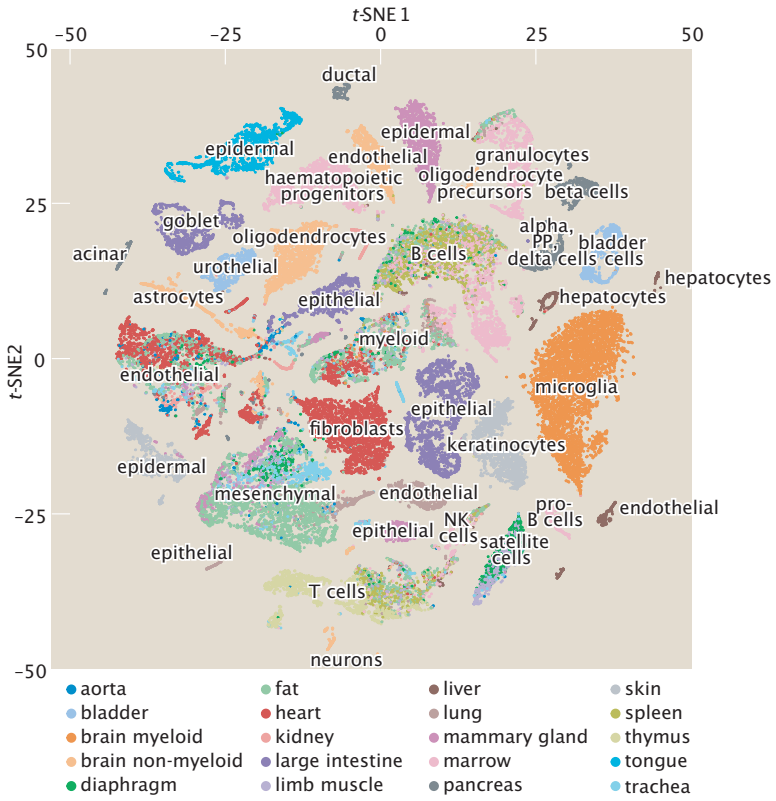

Figure 3: Graphical representation of the *Tabula muris* single-cell sequencing data. Individual cells of different organs in the mouse were subjected to single-cell transcriptome sequencing. Each dot represents a single cell, with its high-dimensional gene expression vector reduced to a two t-SNE lower dimensional representation. Clustering and manual annotation reveal different tissues and cell types. Adapted from The Tabula Muris Consortium et al., *Nature* 562:367-372, 2018.

sents. Further, identify the gene expression vector that corresponds to the number of mRNA molecules detected for all species in cell number 1.

(B) To begin to get a feeling for this kind of data, we imagine an experiment on cells containing only two genes. These cells can adopt three different fates based on the expression state of these genes (i.e., low/low, low/high and high/high). Further, let's assume that these two genes are constitutively expressed, and that low and high gene expression levels correspond to an average of 10 and 35 mRNA molecules per cell, respectively. To remind ourselves of what the null hypothesis for constitutive promoters looks like, write the chemical master equation for a constitutive promoter and show that solving this equation in steady state results in a Poisson distribution. In the case of the low and high expression levels, give the formula for the specific Poisson distribution for those two cases.

(C) Plot histograms of the number of mRNA molecules of gene 1 and gene 2 for each cell type, assuming 1,000 cells of each type. This means that you will invoke the Poisson distribution you derived in the previous part of the problem and use it to describe the distribution of mRNA counts for the different cell types.

(D) Generate a synthetic transcriptome matrix $\mathbf{G}$ with 1,000 cells of each type (for a total of 3,000 cells in your dataset) by sampling from the Poisson distributions that you derived above. Make a plot of this low-dimensional synthetic transcriptome data set consisting of number of mRNA molecules of gene 2 vs. number of mRNA molecules of gene 1, where each dot within the plot corresponds to an individual cell.

Now, we will imagine that we are given this transcriptome data without any more information than the fact that there should be three cell types within it. Note that in reality we will rarely have information about number of cell types within a sample a priori. However, this is a good first step toward building intuition about the challenges of analyzing single-cell sequencing data.

In order to find cell types in our synthetic transcriptome, we will resort to so-called k-means clustering. The steps of this algorithm are illustrated in figure 4 and can be enumerated as follows:

1. The transcriptome data is plotted. In this case, because we only have two genes, this corresponds to a two dimensional plot of the number

of mRNA molecules of gene 2 as a function of the number of mRNA molecules of gene 1 for each single cell. A set of N random points within this data set are then selected, with N being the number of clusters we are trying to identify. These N points will be called the centroids.

2. The distance of every data point to the centroids is calculated. Each data point is assigned to its closest centroid. This is our first approximation to the assignment of cells to our three clusters.

3. Based on the categorization of data points, new centroids are calculated. For each cluster, calculate their corresponding centroids by taking the average values of expression for the two genes.

4. Data points are reassigned to their closest centroid. This means that we now need to take every data point and compute the distance to all three updated centroids and then to assign them to the centroid they are closest to.

5. Steps (3) and (4) are repeated until convergence is achieved.

(E) Write a k-means algorithm to find 3 clusters in your synthetic transcriptome data set. In doing so, generate intermediate plots for the iterations of the algorithm such as those shown in Figure 4.

(F) One of the biggest drawbacks of k-means clustering is that we need to commit to a given number of clusters in advance. Explore what happens if you tell your algorithm to look for two and four clusters instead of three. Document some of the final answers from the algorithm and comment on why it converged to that answer. Comment on how all of these answers correspond to what you actually know about the system given that you generated the transcriptomes!

Finally, it is important to note that all algorithms are limited in the sense that they require commitments by specifying parameters. In k-means, we had to commit to a number of clusters. However, there are other approaches to finding clusters that do not require specifying cluster number a priori such as DBSCAN.

(G) Read about DBSCAN and explain how it works by drawing a graphical example (this can be in cartoon form). For this algorithm, what are the parameters we need to commit to?

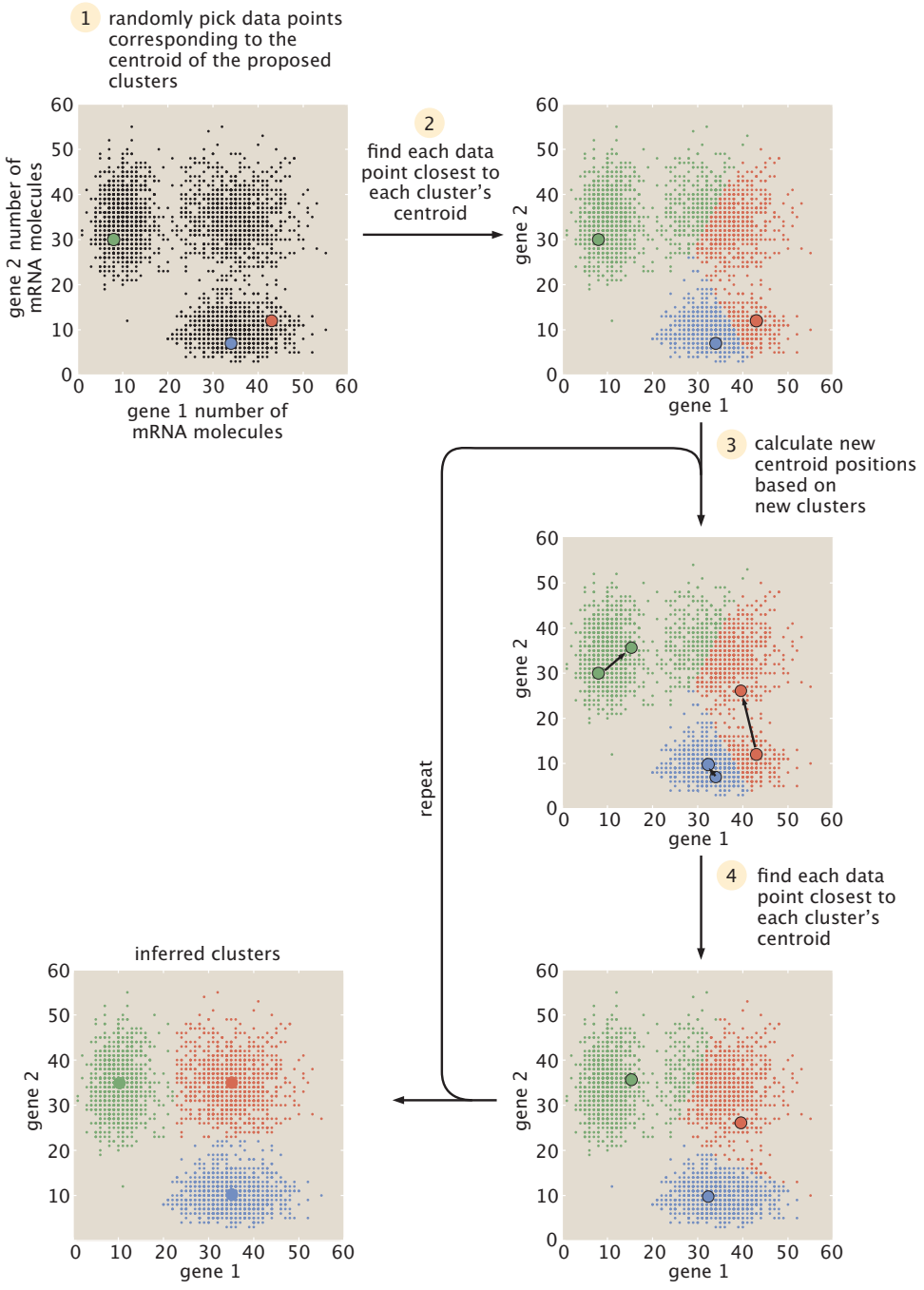

Figure 4: The k-means clustering algorithm. (i) A set of N points are chosen randomly from the dataset to become the centroids of the N clusters to identify. (ii) Each data point is assigned to its closest centroid. (iii) New centroids are calculated for each new cluster. (iv) Data points are reassigned to their new centroids. By iteratively repeating steps (iii) and (iv) convergence can be ultimately reached.