MCB137L/237L: Physical Biology of the Cell
Spring 2025
Homework 8
(Due 3/18/25 at 2:00pm)

Hernan G. Garcia

"How can the events in *space and time* which take place within the spatial boundary of a living organism be accounted for by physics and chemistry?" - Erwin Schrödinger **What is Life?**

1 Finding the Right Coordinates in a Simple Gene Regulatory Network

In the previous homework you synthesized a transcriptome of three genes and three cell types. The expression of each gene for each cell type was described by a Poisson distribution with means given by

$$\text{cell type } 1 = [100, 100, 1], \tag{1}$$

$$\text{cell type } 2 = [100, 1, 100] \tag{2}$$

and

$$\text{cell type } 3 = [1, 100, 100], \tag{3}$$

where $[m_1, m_2, m_3]$ corresponds to the mean levels of genes 1, 2 and 3, respectively.

Further, you performed PCA on this synthetic transcriptome to find the natural coordinate system that describes the synthetic data, and found that you can engage in dimensionality reduction by not considering the third principal component. In this problem, we explore how PCA can be used to analyze more complex data. Specifically, we will generate a transcriptome with extra genes that are activated or repressed by our original three "master genes" shown in Figure 1(A).

We will start with the transcriptome you put together for the previous homework and add 30 genes to it. Figure 1(B) shows the proposed regulatory relations between these 30 new genes and our three master genes. Specifically, each master gene will drive 10 genes, five of them as an activator (leading to an average expression level of 100 mRNA molecules) and

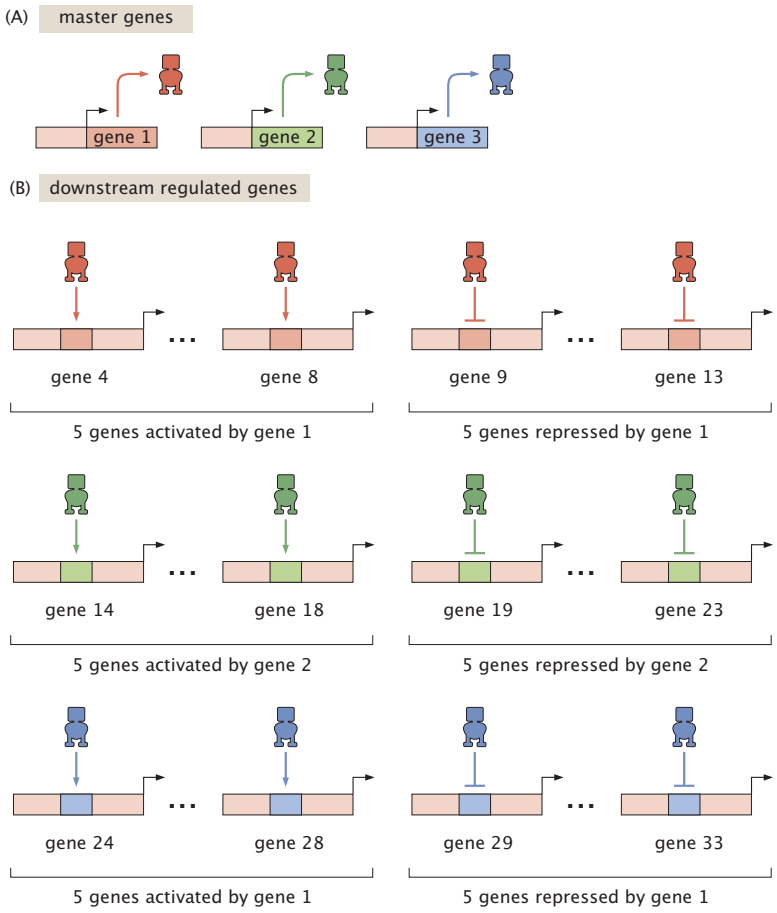

Figure 1: A simple synthetic transcriptional network. (A) Three master genes drive downstream gene expression. (B) These master genes act as activators and repressors on different genes in the network.

five of them as a repressor (leading to an average expression level of 1 mRNA molecule).

(a) Make a heatmap of your new gene expression matrix **G** reporting on the number of mRNA molecules for each gene in each cell. Make sure to include a colorbar indicating what number of mRNA molecules each color corresponds to.

(b) Calculate the covariance matrix of your transcriptome and find its eigenvectors and eigenvalues. Sort your eigenvalues according to size, and plot them.

(c) Comment on why two of your eigenvalues are larger than the rest. Then, project your data onto the two eigenvectors corresponding to these eigenvalues and plot it.

Now, we calculate how much error you made by adopting this reduced dimensionality description. To make this calculation, project your data onto the new coordinate system defined by your PCA. For example, if we define the distance along each PCA i for cell j as $d_{j,i}$, the

error in considering only the first two principal components for cell j is

$$\text{error}_j = \sqrt{\frac{\sum_{i=3}^{30}(d_{j,i})^2}{(d_{j,1})^2 + (d_{j,2})^2}}. \tag{4}$$

(d) For each cell, calculate the error incurred in only considering the first principal component and plot this error in a histogram. Repeat this calculation and histograms for the case where you consider only the first two and the first three principal components. How do you justify taking only the first two principal components as a reasonable reduced description of our system?

2 Finding the Right Coordinates in a The Presence of Uninformative Genes

In this problem, we turn our attention to the challenge of finding the right coordinates of our transcriptome when our three original genes are embedded in a transcriptome with several genes whose expression is uncorrelated with cell type. Specifically, we will define a new transcriptome. As shown in Figure 2(A), this new transcriptome still contains the three master genes dictating cell type that we have considered so far. In addition, the transcriptome will also contain 10 genes that do not inform the cell fate decision. Instead, as shown in Figure 2(B), *for each cell and each gene*, a coin is flipped in order to determine whether the gene will be expressed at a high or low level.

(a) Make a heatmap of your new gene expression matrix **G** reporting on the number of mRNA molecules for each gene in each cell. Make sure to include a colorbar indicating what number of mRNA molecules each color corresponds to.

(b) Calculate the covariance matrix of your transcriptome and find its eigenvectors and eigenvalues. Sort your eigenvalues according to size, and plot them.

(c) Project your data onto the two eigenvectors corresponding to the two largest eigenvalues and plot it.

(d) For each cell, calculate the error incurred in only considering the first principal component and plot this error in a histogram. Repeat this calculation and histograms for the case where you consider only the first two and the first three principal components.

(e) Compare the results from your projection in (c) to the error incurred in this projection calculated in (d). How do you justify taking only the first two principal components as a reasonable reduced description of our system? Specifically, comment on your ability to determine that there are three cell types versus your ability to quantitatively reproduce gene expression states using your reduced dimensionality description.

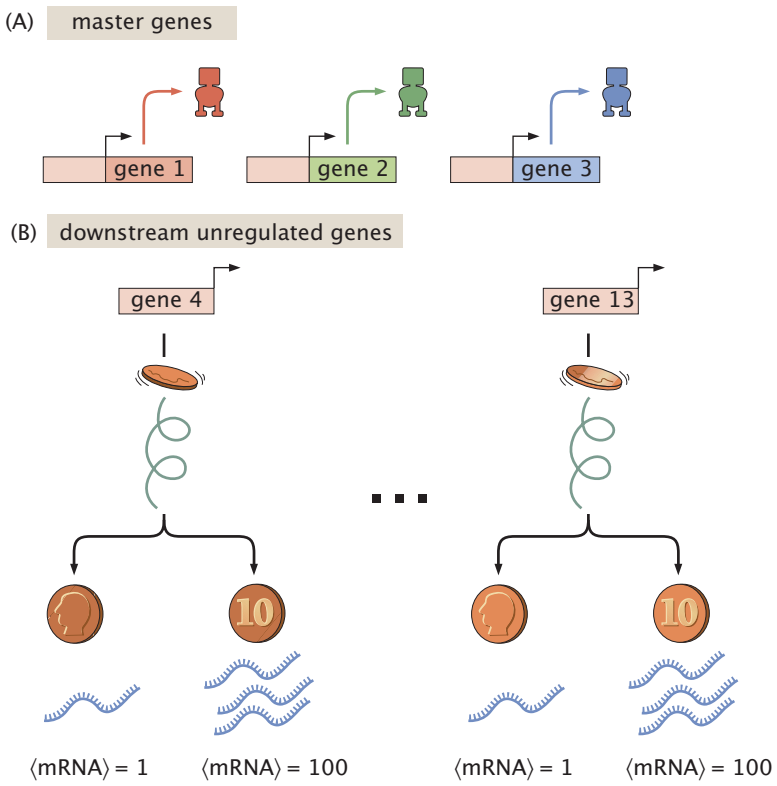

Figure 2: A synthetic transcriptome with uninformative genes. (A) Three master genes determine cell type. (B) The transcriptome also contains 10 extra genes uncorrelated with cell type. For each cell and gene, an honest coin is flipped to determined whether that gene will be expressed at a high or low level.