Bi 1: The Great Ideas of Biology
Homework 6
Due Date: Thursday, May 18, 2023

> "It is particularly incumbent on those who never change their opinion to be secure of judging properly at first."
>
> Jane Austen

Gene Regulation in Modern Biology

Though the central dogma of molecular biology teaches us about the chain of molecular processes connecting DNA to the functional proteins that run a cell, it says nothing about when and where the genes of a given organism will be turned on. One of the most important discoveries in the emerging field of molecular biology was made by the French scientists Jacques Monod and Francois Jacob, both participants in the dramatic events of World War II, who shortly after the end of that great conflict discovered how genes are regulated. Despite the fact that we have learned much about how genes are regulated that goes beyond their original insights, much of the picture they formulated has stood the test of time. In this problem, we begin our own detailed investigation of how the proteins known as transcription factors can tune the level of gene expression to correspond to the demands of an organism.

Fig. 1 shows the processes of transcription and translation, mediated by RNA polymerase and the ribosome, respectively. But we also see that the process of transcription can be up- and down-regulated through the action of activators and repressors. Note that the repressor can bind to the promoter region of the DNA and thereby inhibit the ability of the RNA polymerase to bind, thus reducing transcription. In the remainder of this problem and next week's problem as well, we work out both the theoretical description of this process as well as the analysis of the kinds of experiments using fluorescent proteins that allow us to precisely measure gene expression.

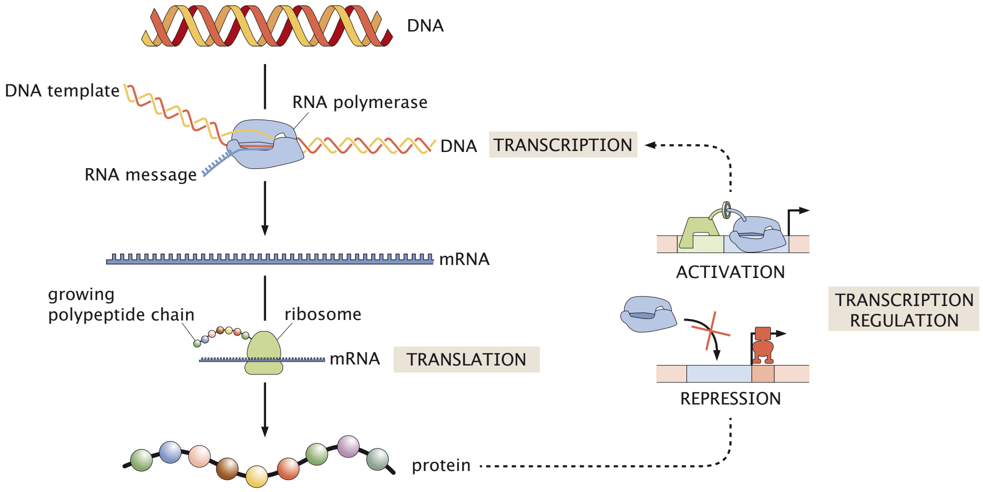

Figure 1: **The role of transcriptional regulation in the central dogma.** The central dogma involves the synthesis of mRNA through transcription and the synthesis of protein in the process of translation. Here we note that transcription is controlled by activators and repressors that turn genes on and off, respectively.

1. Statistical Mechanics of an unregulated promoter

In this homework we will walk through the complete details behind the derivation of the fold-change equation sketched in class. First we will be thinking about an unregulated promoter, i.e. a promoter that is not controlled by any kind of transcriptional activator or repressor.

You might recall that with our statistical mechanics protocol we imagined the entirety of the genome to be built up of N_{NS} non-specific binding sites. Experimental evidence shows that even when the RNA-polymerase (RNAP) is not transcribing, it spends most of its time bound to the genome at either specific or non-specific binding sites. That is why we ignored the possibility of having polymerases in solution and we counted how many ways we can arrange P RNAP molecules on the genome, assigning specific energies ε_{pd}^{S} and ε_{pd}^{NS} for the state in which a polymerase is bound to a specific or a non-specific binding site, respectively.

Question 1a

Fill out a table like that shown in Fig. 2 by writing down the energy terms, the multiplicity and the Boltzmann weight for the two states shown. Though the schematic figure only shows four polymerases (in blue), assume there are P of them distributed throughout the genome.

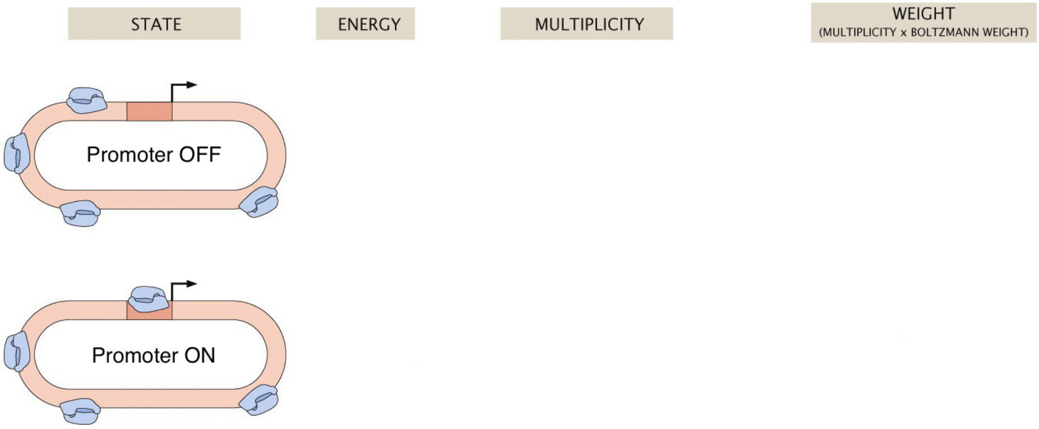

Figure 2: **States and weights for an unregulated promoter.** The circular bacterial genome is represented as a light orange oval and the promoter region is shown in darker orange. Assume there are a total of P RNA polymerases bound on the genome.

The main assumption of thermodynamic models of gene regulation can be expressed as

$$\text{gene expression} \propto p_{\text{bound}}. \tag{1}$$

In words, this means that the expression level of a gene is proportional to how likely it is to find the RNAP bound to the promoter, which we denote by p_{bound}. In principle to remove the proportionality sign we would have to know something about the rate of expression, but as we will show later by computing the fold-change, one can get around that caveat provided that the regulatory proteins do not affect the kinetics of the transcription process itself.

Question 1b

Given the states and weights you developed in the previous question, compute p_{bound}, i.e. the probability of finding an RNAP bound to the promoter.

(Hint: Recall that Boltzmann's law tells us that the probability of a microstate with energy E is of the form $P(E) \propto e^{-E/k_B T}$ where k_B is the Boltzmann constant, and T is the temperature. Recall also the approximation we have used in class multiple times where $\frac{N_{ns}!}{(N_{NS}-R)!} \approx (N_{NS})^R$ for $N_{ns} \gg R$)

Question 1c

Simplify the previous equation by multiplying and dividing by the inverse of the weight of the empty promoter state, and defining $\Delta\varepsilon_p = \varepsilon_{pd}^S - \varepsilon_{pd}^{NS}$. Show that your previous results can be simplified to the form

$$p_{\text{bound}} = \frac{\frac{P}{N_{NS}} e^{-\Delta\varepsilon_p/k_B T}}{\frac{P}{N_{NS}} e^{-\Delta\varepsilon_p/k_B T} + 1}. \tag{2}$$

2. Linear models of genotype-to-phenotype maps

The RNA polymerase binding site is ≈ 40 basepairs long. What that means is that the binding affinity of the polymerase is encoded in the sequence of these 40 basepairs. A current active research direction in biology is the development of quantitative models that can map from a specific sequence to a meaningful biophysical parameter such as the binding energy. Doing this from first principles remains for now intractable given the complexity of the molecular players involved. But with the use of modern DNA sequencing along with fluorescent gene expression reporters, scientists have developed methods to infer effective binding energies.

The simplest model to map from sequence to binding energy is to assume each basepair in the binding site contributes independently to the total binding energy of the RNA polymerase to DNA. This implies that for a given sequence of length l the total binding energy is given by

$$\Delta\varepsilon(\mathbf{L}) = \sum_{i=1}^{l} \varepsilon_i, \tag{3}$$

where ε_i is the energy associated with whether the i^{th} base is an A, C, G or T. What Eq. (3) is telling us is that in order to compute the binding energy for the sequence $\mathbf{L}$ we must go through each of the l positions, look at the identity of the basepair and add the corresponding energy contribution for the base pair at position i. These linear models have proven to be very powerful in identifying the relative contributions of each position to the affinity of a DNA-binding molecule. One such map is shown in Fig. 3. This map for the RNAP was used to design promoters with different strengths. To give you an intuition for how this works: if you look at position -34 you can see that letters A and C are colored deep red, while G is in dark blue. The numerical entries of the matrix for this position are:

1. A: 1.500683 k_BT
2. C: 1.490967 k_BT
3. G: -0.313427 k_BT
4. T: 0.633869 k_BT

which means that if the promoter sequence were to have an A at position -34 it would contribute positively to the binding energy, i.e. it would reduce the affinity of the RNAP for the promoter. The preferred basepair at this position would be a G, which would contribute negatively to the binding energy, implying that the binding energy is favorable in this case. For this problem you will use this map to compute the binding energy for different promoter sequences.

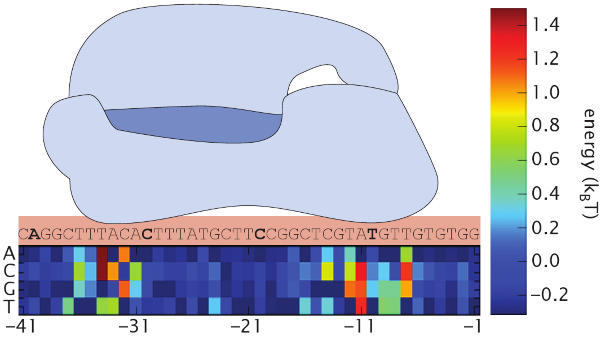

Figure 3: **Sequence-to-energy map for the RNAP.** The $l \times 4$ matrix maps from a given nucleotide (A, C, G or T) to a corresponding energy contribution ε_i in units of k_BT. From the color map we can see that there are two blocks of basepairs at positions -10 and -35 that contribute the most to the binding energy of the RNAP (positions are indicated with respect to the transcription initiation site).

Question 2a

From the course website next to where you downloaded this problem set you will find a link to a supplemental file that contains the numerical values of the energy matrix shown in Fig. 3. The matrix covers base pairs [-41:-1] where 0 denotes the transcription start site. In this file, each row corresponds to a given position, and each column corresponds to a value for that base pair. The columns are ordered [A,C,G,T]. Using this matrix, compute the binding energy for the following promoters:

1. `p1 : TCGAGTTTACACTTTATGCTTCCGGCTCGTATAATGTGTGG`
2. `p2 : TCGAGTTTACACTTTATGCTTCCGGCTCGGATAATGTGTGG`

To make things easier we provide you with the wild-type *lac* promoter sequence `CAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGG` that has a reported binding energy of -5.35 k_BT. You therefore need only to find the difference between the wild-type sequences and the two sequences listed above, then compute the energy difference using the energy matrix.

(Hint: It might be useful to open the text file in Python to look for the basepair changes. You could also use an online alignment tool such as this one. Just make sure to modify the gap open penalty to a much higher value. What this means is that we highly penalize having any gap between the sequences, since we want a basepair by basepair comparison).

Question 2b (extra credit)

Write a Python function that takes as an input a promoter sequence and an energy matrix and computes the binding energy automatically. Use it to compute the binding energy for the following sequences:

1. `CAGGCTCTACGCTTTATTCTTGCGGCTCGTATGGTGTGTGG`
2. `GAGGCTGGACACTTTAATCTTCCCTATCGTATGTTGTGTGC`
3. `CGAGCATGCCACCTTAAGCCTCTGGCTCGTATGACGTGTGG`

You may paste a screenshot of your code, along with the results, to your submission.

3. The "simple repression" architecture

One of the most common regulatory architectures in *E. coli* is the so-called simple repression motif. As we saw in class, this consists of a single binding site for a transcriptional repressor. In this problem we will derive the fold-change equation that describes the relative expression level of a gene due to the presence of a repressor.

Question 3a

Fill out the table shown in Fig. 4. This time you can just write the renormalized weights, i.e. the Boltzmann weights multiplied by the inverse of the empty promoter weight. What this means is that for Question 1a, we had that the empty promoter weight was given by

$$w_{empty} = \frac{(N_{NS})^P}{P!} e^{-P\varepsilon_{pd}^{NS}/k_BT}. \tag{4}$$

But if we normalize by multiplying by the inverse of this weight we should end up with $w_{empty} = 1$.

If you still feel you need to go through the algebra, that's fine! Feel free to explain each of the steps you took to simplify the weights.

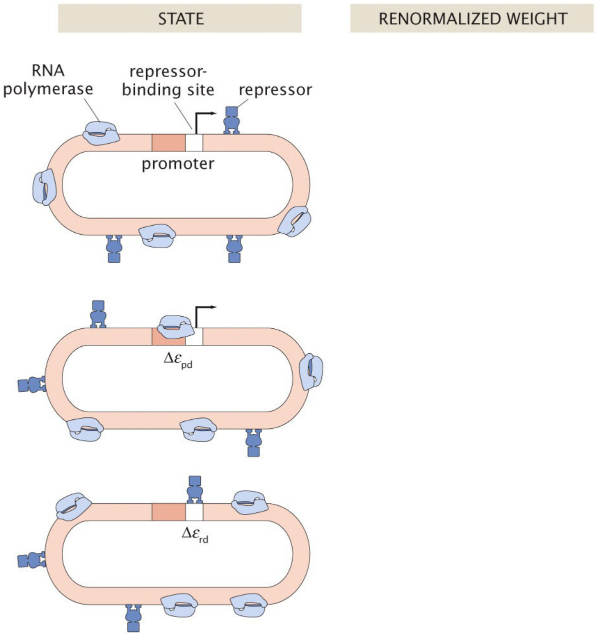

Figure 4: **States and weights for the case of simple repression.**

Having computed the weights for each of the states, let's compute p_{bound} for the simple repression architecture.

Question 3b

Compute p_{bound}, i.e. the probability of the RNAP being bound to the promoter, taking into account the new state with the repressor.

Given the weights for the different states, we can compute the probability for each of the states. Let's explore now how probable each of the states is given some literature-based values of the parameters.

Question 3c

Given the parameters shown in Table 1, plot the probability of each of the 3 states as a function of the number of repressors per cell R. Let R range from 1 to 1000, and make sure to show the x-axis in log scale. (If you choose to do this in Python, the functions `np.logspace` and `plt.xscale` may be helpful).

Table 1: Physiologically-relevant values for each parameter.

Parameter	**Value**	**Units**
P	1000	RNAP/cell
$\Delta\varepsilon_p$	-5	k_BT
$\Delta\varepsilon_r$	-15	k_BT
N_{NS}	4.6×10^6	basepairs/genome

Having calculated p_{bound} in the presence and in the absence of the transcription factor we can now proceed to compute the fold-change in gene expression. But before going into that let's remember why we compute this quantity rather than the absolute level of gene expression itself. As we saw in class the simplest model for gene expression is given by the differential equation

$$\frac{dm}{dt} = r \cdot p_{\text{bound}} - \gamma m, \tag{5}$$

where m is the average mRNA copy number, r is the mRNA production rate and γ is the mRNA degradation rate. Experimentally what we measure is a single snapshot of cells in steady-state. Furthermore what we measure in our experiments is not the amount of mRNA but the the amount of fluorescence given by a protein encoded in the mRNA. About this last point people have spent a lot of time figuring out how to measure protein and mRNA with different readouts; what they have found is that at least at the mean level there is a rock-solid linear relationship between these two. That means that our naive model in equation Eq. (5) is a good approximation.

If fold-change is defined as

$$\text{fold-change} \equiv \frac{\text{gene expression}_{ss}(R > 0)}{\text{gene expression}_{ss}(R = 0)} = \frac{m_{ss}(R > 0)}{m_{ss}(R = 0)} \tag{6}$$

where the subscript ss denotes the steady state, we must then compute this steady state mRNA expression level in the presence and in the absense of the repressor.

Question 3d

Compute the steady state of Eq. (5) and solve for m, then compute the fold-change in gene expression due to the presence of the repressor. Note from Eq. (5) that only one of the terms depends on the repressor copy number.

Now that we've justified the use of fold-change versus gene expression level directly, let's get the final expression.

Question 3e

Compute the fold-change in gene expression given the equations you derived for p_{bound} with and without the repressor. Then, given the parameter values shown in Table 1, you should be able to drop one of the weights given its relative magnitude compared to the others. When the dust settles you should end with an expression of the form

$$\text{fold-change} = \frac{1}{1 + \frac{R}{N_{NS}} e^{-\Delta\varepsilon_r / k_B T}} \tag{7}$$

Finally, what does this equation predict?

Question 3f

On a log-log plot, show the fold-change as a function of the repressor copy number for a weak ($\Delta\varepsilon_r = -10\ k_B T$), a medium strength ($\Delta\varepsilon_r = -14\ k_B T$), and a strong repressor binding site ($\Delta\varepsilon_r = -17\ k_B T$). This theoretical prediction is the basis of the experiments you will interpret in the next assignment!