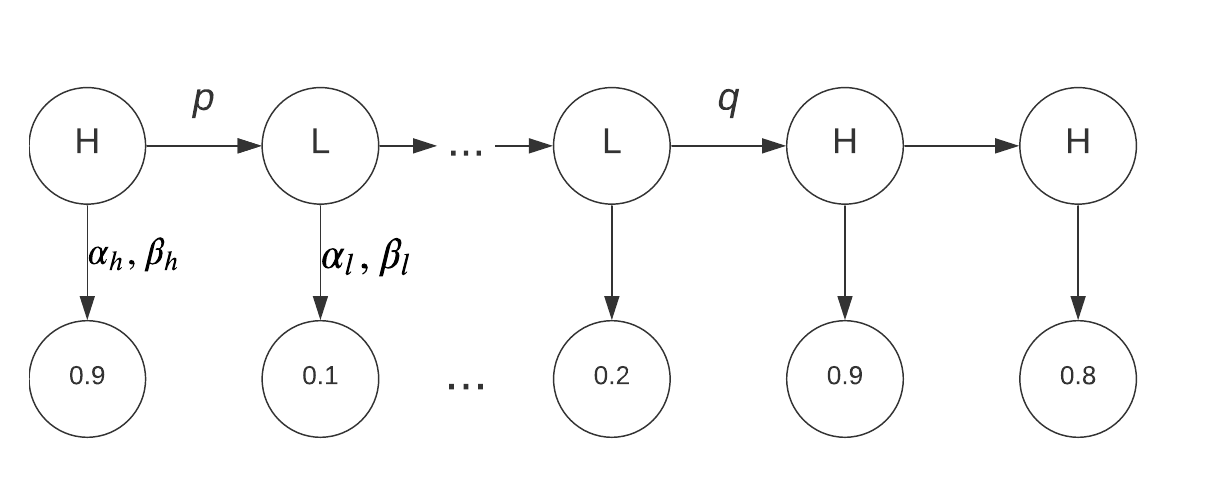
Using the Beta Distribution to Model DNA Methylation Regions

Research Proposal

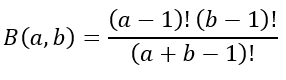
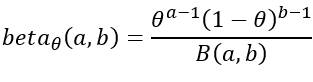
Yuval Dellus, Liron Gershuny, Yonatan Chamudot, Yoni Rosensky, Yacov Averbuch

**Research Question:**   
 Can the distribution of methylated CpG sites be modeled using a Hidden Markov Model with emission probabilities described by beta distribution parameters?

We propose a model that divides the genome between high-methylation regions (within CpG islands) and low-methylation regions, where the methylation regions serve as the hidden states in a Hidden Markov Model (HMM). At each CpG site within the region, we will count the observed instances in which the site is methylated across samples of the same tissue. These counts will serve as the observed states in our HMM. Using the Baum-Welch algorithm we saw in class, our model will learn the transition probabilities between states and the emission probabilities within each state.



Instead of using static emission probabilities for each hidden state, we will attempt to describe the emission probabilities using the beta distribution. For each type of methylation regions our model will learn and parameters that define a beta distribution for that region. Using the learned beta distribution, we will calculate the likelihood of observing the methylated sites in that region for the Expectation stage. The algorithm then continues to learn transition and emission probabilities until our model converges.

The PDF of a beta distribution can be represented as:  


, Where

We chose to base our emissions model on the beta distribution because of its flexibility, since its parameters can account for different means and variances of the distribution. This is a very powerful property for training our model on different tissues and different methylation patterns.

Our data will consist of methylation data for CpG sites in a given tissue. The data is formatted as an array of tuples, where each tuple represents a CpG site. Each entry in the data contains multiple observations of DNA sequences in the tissue: the second entry in each tuple represents the number of successful reads of the CpG site, and the first entry represents the number of times the site was methylated.



One challenge in constructing this model lies in calculating the likelihood of a set of discrete observed values given a continuous beta distribution. We will attempt to solve this problem by calculating the integral at a small range around the discrete points:

where is the CDF of the beta distribution with parameters and .

Another challenge lies in choosing initial transmission and emission probabilities for our model. We will initialize our model by making an initial pass over the data and converting it to an array of percentages. We then pass the array through a smoothing function to reduce the effects of sharp drops on the initial transmission probabilities. The entries in the array are converted to High or Low states based on a threshold of 0.5, and finally we count the number of transitions between states in order to derive the initial transmission probabilities. Lastly, we run a scipy beta distribution function over the statistics for each hidden state in order to calculate the initial emission parameters.