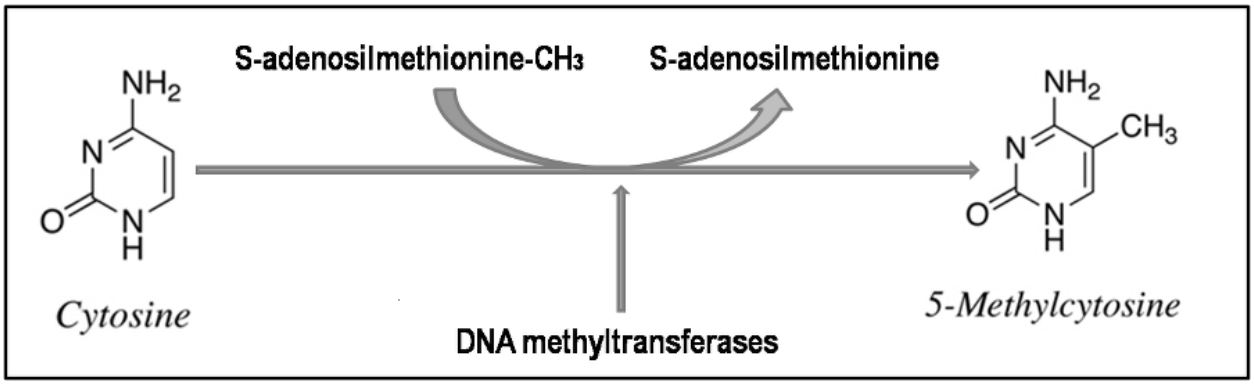
A Continuous Output Hidden Markov Model for DNA Methylation Percent at CpG Sites

Yacov Averbuch, Yonatan Chamudot, Yuval Dellus, Lion Gershuny, Jonathan Rosenski

In this work we will create a continuous output HMM for DNA methylation percent (methylation beta value) using k hidden states where the output of the methylation percent (or number of methylated reads vs total reads) is modeled by a binomial distribution

Biological background

Methylation, chemically, is the transfer of a methyl group (-CH3) to an organic compound. These methyl groups can be transferred through addition and substitution reactions where the methyl group takes the place of a hydrogen atom on the compound.
In the field of biology, DNA methylation is an epigenetic mechanism where there is an addition of a methyl group to a cytosine residue, often to the fifth carbon atom of a cytosine ring, causing cytosine to become 5-methylcytosine and used by cells in order to control gene expression or activity. This conversion of cytosine bases to 5-methylcytosine is catalyzed by DNA methyltransferases.



DNA methylation occurs at CpG sites—that is, sites where a cytosine lies next to a guanine base and the result is two-methylated cytosines positioned diagonally to each other on opposite strands of DNA.

Tumors begin with abnormal localized hypermethylation, genome-wide hypomethylation, and increased expression of DNA methyltransferase. Research shows that genome-wide hypomethylation leads to increased mutation rates and instability of chromosomes.
Bacteria use methylation as a tool for self-defense. Bacterial cells protect their DNA through the methylation of adenine or cytosine bases. Foreign DNA that enters the bacteria remains unmethylated and therefore is prone to destruction by the bacteria's restriction enzymes.
DNA methylation is an epigenetic event that is involved in embryonic development and cell cycle regulation; hence analyzing DNA methylation of any cell provides valuable information about its cellular state, its developmental potential, and its overall health.
CpG islands are defined as stretches of DNA 500–1500 bp long with a CG: GC ratio of more than 0.6 compared to the rest of the genome. Methylation is sparse but global in mammals, CpG sequences accounts for about 1–2% across the genome, aside from certain stretches (of around one kilobase) where the content of CpG is high – those are CpG islands.
CpG islands typically reside at the 5' ends of genes, and the majority of all human genes have CpG islands at their 5' end. However, CpG islands also can be found in close proximity to transcription start sites and can induce gene expression by preventing the binding of insulators that repress gene expression, suggesting there is an established recognition system. When CpG islands become aberrantly hypermethylated, it is generally associated with decreased expression of the gene by preventing the binding of factors to the DNA that promotes transcriptional activity.

While overall methylation levels and completeness of methylation of particular promoters are similar in individual humans, there are significant differences in overall and specific methylation levels between different tissue types and between normal cells and cancer cells from the same tissue. Proteins that bind to methylated DNA also form complexes with the proteins involved in deacetylation of histones. Therefore, when DNA is methylated, nearby histones are deacetylated, resulting in compounded inhibitory effects on transcription. Likewise, demethylated DNA does not attract deacetylating enzymes to the histones, allowing them to remain acetylated and more mobile, thus promoting transcription.

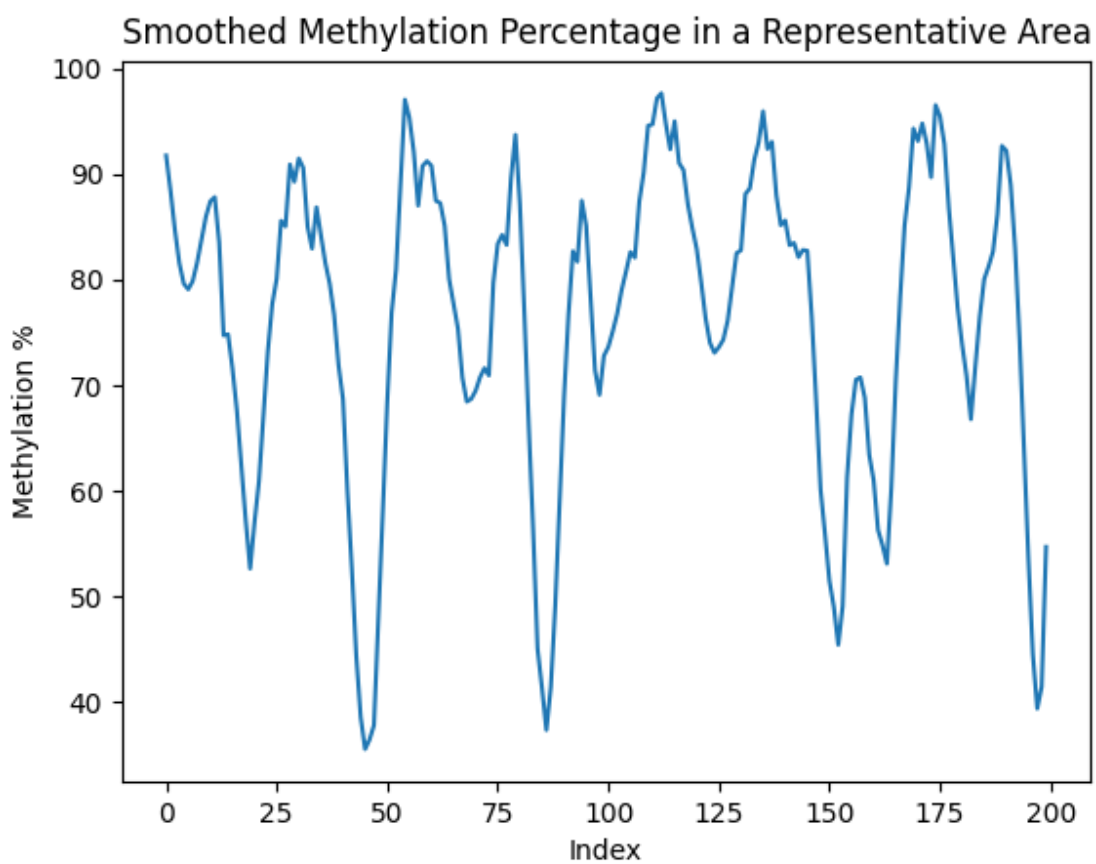
Given the critical role of DNA methylation in gene expression and cell differentiation, it seems obvious that errors in methylation could give rise to a number of devastating consequences, including various diseases. A large amount of research on DNA methylation and disease has focused on cancer and tumor suppressor genes. Tumor suppressor genes are often silenced in cancer cells due to hypermethylation. In contrast, the genomes of cancer cells have been shown to be hypomethylated overall when compared to normal cells, with the exception of hypermethylation events at genes involved in cell cycle regulation, tumor cell invasion, DNA repair, and other events in which silencing propagates metastasis. In fact, in certain cancers hypermethylation is detectable early and might serve as a biomarker for the disease.

It is known that DNA methylation is normally "off" or "on" for extended ranges of continuous CpG sites. This means that CpG sites normally are methylated or unmethylated depending on what the previous CpG site state is, unless that CpG site switches from the "off" mode to the "on" mode. For example it is common to see such scenarios: CCCCCCCCTTTTTTTTCCCCCCCC, where a C represents a methylated CpG site and a T represents an unmethylated CpG site.

Motivation for our Model

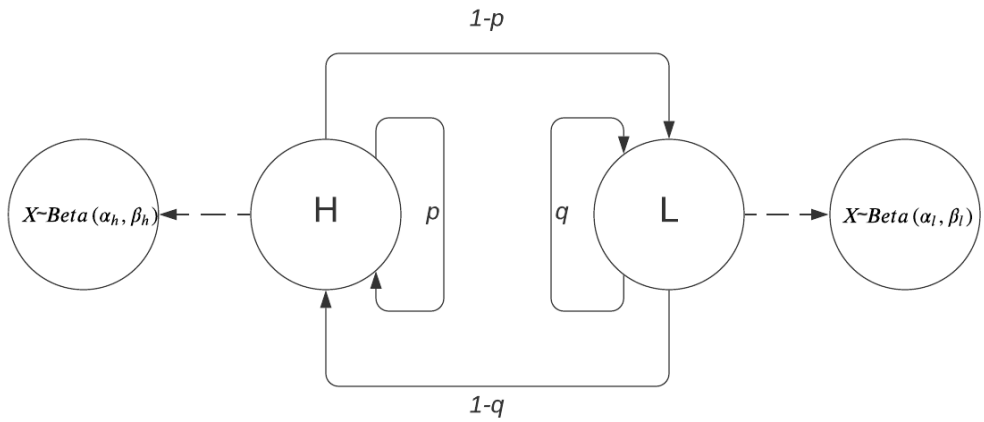
We began by looking at Whole Genome Bisulfite Sequencing (WGBS) data from Tommy Kaplan's Lab. We focused on one specific sample of prostate epithelial cells. The WGBS data was formatted as a beta file, which presents three data points for every CpG site in the sample. The first data point is the CpG index according to an hg19 reference genome. The second data point is the number of reads overlapping this site in which the site was found to be methylated. The final data point is the total number of reads which included this CpG site. We processed this raw data to produce an array that represented the percentage of reads in which the site was methylated (methylated reads/total reads).

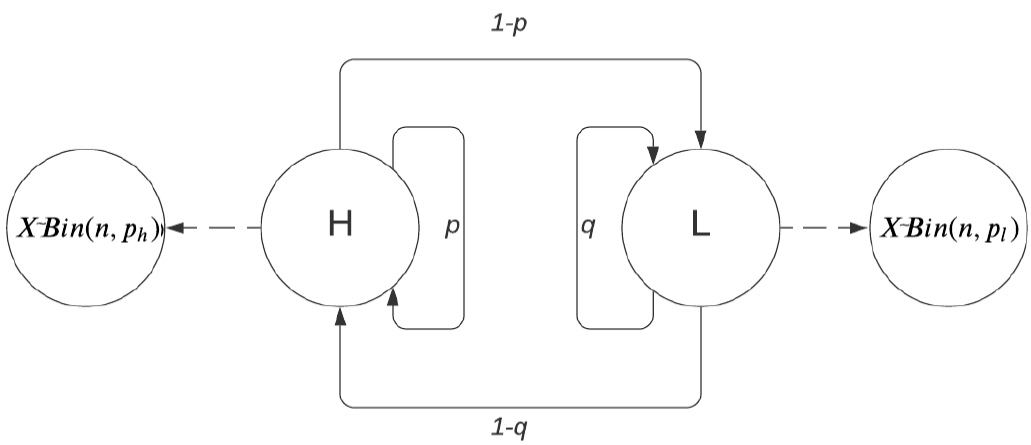
In studying the methylation data at various areas in the sample, we noticed that the percentage of methylated samples at each position tended to jump between high percentages and low ones, with sharp transitions between high and low "streaks". This led us to propose a model where the genome could be divided into two types of regions: "high" methylation regions, where each CG site would, with relatively high probability, have a high percentage of methylated samples, and "low" methylation regions, where each CG site would, with relatively high probability, have a low percentage.



HMM

In keeping with our theory of “high” and “low” methylation regions, we initially proposed a Hidden Markov Model with two hidden states. The “High” hidden state will emit methylation percentages according to a beta distribution with a mean (peak) located towards the higher end of the distribution. In other words, when the model is in a high methylation region it will output higher methylation percentages on average. Conversely, the “Low” hidden state will emit methylation percentages according to a beta distribution with a mean located towards the lower end of the distribution. We will denote the parameters of the “high” beta distribution as α_h, β_h and the “low” beta distribution as α_l, β_l. A “high” state repeats with probability p and a “low” state repeats with probability q. The theorized HMM can be represented by the following state machine diagram:





Each CpG site is an element in the input sequence. Thus transitioning from H to L means that for some CpG site index **j** the hidden state is H and for CpG site index **j+1** the hidden state is L. Here each hidden state (H or L) emits observations from the labelled distribution.

The learning algorithm:

Once the state machine is defined we will use the Baum-Welch algorithm (a version of the expectation-maximization algorithm learned in class) to learn the emission probabilities of each hidden state as well as transition probabilities between hidden states for adjoining CpG sites. That is, the probability to see **m** methylated observations at site **i** and **n-m** unmethylated observations, given the hidden state at site **i-1**, is independent of all previous sites.

Here we will not describe the Baum-Welch algorithm in detail (please see lecture notes for details) but rather discuss what each part of the algorithm represents in the context of CpG methylation.

The input to the algorithm is a list of sequences. Each sequence is composed of **n** CpG sites. In each index of a sequence is the number of methylated observations and the number of unmethylated observations. Using this we can calculate the sufficient statistics necessary for a binomial distribution. In our definition of the state machine we have two hidden states: high and low. Each one has unique parameters to a binomial distribution. At each iteration of the EM algorithm we first estimate the expected values of the sufficient statistics and then using those we calculate the maximum likelihood estimates of the parameters of the binomial distributions. If every CpG site was labelled as high or low we would know how to calculate the ML estimates. But we have unlabelled data so we estimate the expected number of methylated observations in high states and the expected number of observations in low states (this generalizes to more than two hidden states). Please see the appendix for the proof/formula for calculating the ML estimations of the parameters given the input data.

After every EM iteration we update the emission probabilities and transition probabilities with those found in the previous iteration. We stop iterating when the log-liklihood stops improving more than a certain amount. We call this parameter for deciding how large gaps in likelihood must be the “convergence threshold”. We usually set the convergence to be 0.1 but we did try a few different values to see how it affected the results.

We encounter a problem with calculating the probability of a one CpG site to be HIGH or LOW using a **beta** distribution: Since the sufficient statistics for a Beta distribution are the number of successes (methylated observations) and the number of failures (unmethylated observations) the larger the genomic region we use, the larger these two numbers are. Even when these numbers are factored by the probability of being hidden state high/low, these numbers are very large. For reasonably sized windows of a 100,000 base pairs, we were seeing our HMM mode estimating the number of methylated/unmethylated observations to be in the thousands. The beta distribution models the probability of seeing a specific success probability of a binomial distribution (it is the conjugate prior of a binomial distribution). Thus the more observations you have, the more confident (or narrow) is the estimated beta distribution. Thus even for these not very large windows, the beta distribution would give a zero probability for any

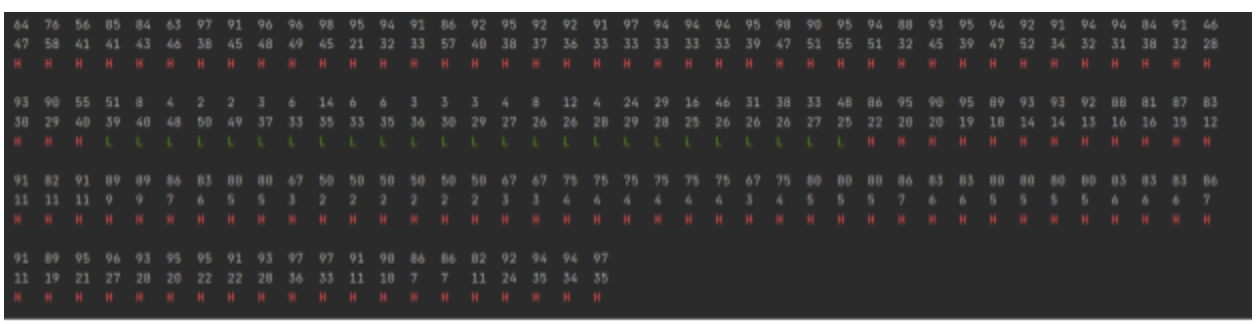
site which was not virtually equal to the mean of the beta distribution. Thus, the HMM modelled with a beta distribution was not able to learn properly.

Initial Findings

To see if our model is able to predict sleep/awake states of methylation we decided to test it on a marker for the Prostate-Epithelial tissue. That is a region, which is unmethylated in Prostate Epithelial tissue and methylated in most other tissues. If you look at the average methylation of each CpG site in this region is fully methylated, then at the location of the marker, becomes unmethylated then fully methylated again.

The average methylation levels across CpG sites 16822953-16823093. Values of 0 imply an average methylation level of between 0-10%, 5 an average methylation level of between 50-60%, and so on. We can see in all the Prostate Epithelial samples one clear region of low methylation in green.

We would like to produce a model which recognizes all of the highly methylated sites as “sleep” and the unmethylated region as awake.

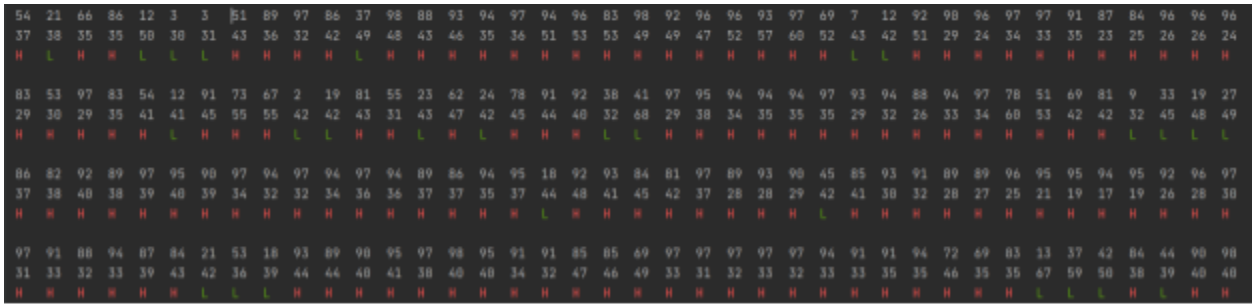


Each column is a CpG site (grouped into rows of 3). The first row is the average methylation level at the CpG site, the second row is the number of observations at that site. The third row is the predicted hidden state (High/Low) according to the maximum posterior probability.

Our HMM was able to classify the region we manually labelled “sleep” as highly methylated (red H) and the region which is unmethylated as “awake” (green L).

It appears that all sites which have low methylation average are marked as green L and all sites with a high methylation average are marked with a red H. Thus, we would like to answer the question: Does our model determine a simple threshold on the methylation average?

So it is easy to see when looking over more CpG sites that our model does not determine a simple threshold:



```
54  21  66  86  12  3   3   51  89  97  86  37  98  88  93  94  97  94  96  83  98  92  96  96  93  97  69  7   12  92  98  96  97  97  91  87  84  96  96  96
37  38  35  35  50  38  31  43  36  32  42  49  48  43  46  35  36  51  53  53  49  49  47  52  57  60  52  43  42  51  29  24  34  33  35  23  25  26  26  24
H   L   H   H   L   L   L   H   H   H   H   L   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   L   L   H   H   H   H   H   H   H   H   H   H   H

83  53  97  83  54  12  91  73  67  2   19  81  55  23  62  24  78  91  92  38  41  97  95  94  94  94  97  93  94  88  94  97  78  51  69  81  9   33  19  27
29  30  29  35  41  41  45  55  55  42  42  43  31  43  47  42  45  44  40  32  68  29  38  34  35  35  35  29  32  26  33  34  60  53  42  42  32  45  48  49
H   H   H   H   H   L   H   H   H   L   L   H   H   L   H   L   H   H   H   L   L   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   L   L   L   L

86  82  92  89  97  95  90  97  94  97  94  97  94  89  86  94  95  18  92  93  84  81  97  89  93  90  45  85  93  91  89  89  96  95  95  94  95  92  96  97
37  38  40  38  39  40  39  34  32  32  34  36  36  37  37  35  37  44  40  41  45  42  37  28  28  29  42  41  30  32  28  27  25  21  19  17  19  26  28  30
H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   L   H   H   H   H   H   H   H   H   L   H   H   H   H   H   H   H   H   H   H   H   H   H

97  91  88  94  87  84  21  53  18  93  89  90  95  97  98  95  91  91  85  85  69  97  97  97  97  97  94  91  91  94  72  69  83  13  37  42  84  44  90  98
31  33  32  33  39  43  42  36  39  44  44  40  41  38  40  40  34  32  47  46  49  33  31  32  33  32  33  33  35  35  46  35  35  67  59  50  38  39  40  40
H   H   H   H   H   H   L   L   L   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   H   L   L   L   H   L   H   H
```

There is one site in the first row where the methylation average is 51 and it is marked at red H and one site which is methylation average 53 in the last row marked as green L.

Our HMM model is more flexible and robust than just predicting high/low methylation based off of a threshold on the methylation average.

The original motivation is that throughout most of the genome there are “on” and “off” continuous regions for which the DNA methylation is either “on” or “off”. If this is the case then we would expect our HMM to learn transition probabilities that high when staying in the same hidden state. Using our algorithm and a convergence threshold of 0.1 on a sequence of CpG sites of lengths approximately 7k CpG sites we get the following transition probability matrix:

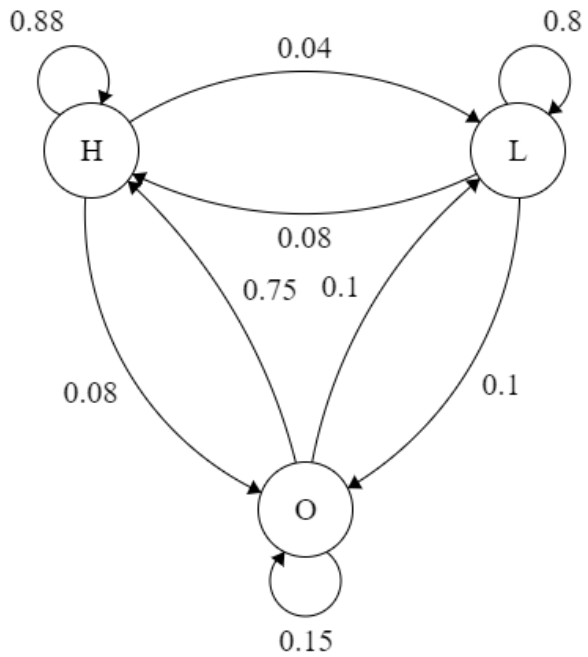
	High	Low
High	0.94	0.06
Low	0.19	0.81

Thus the probability to stay in the same state is significantly higher than the probability to switch states, confirming our initial hypothesis. We did not mention this in our original motivation but during data exploration we did find that there were many more regions of high methylation than low methylation and that the regions of “on” or high methylation tended to be longer than those of “off” or low. This confirms the results found by the HMM which suggests that the probability to go from high to high is significantly higher than from low to low.

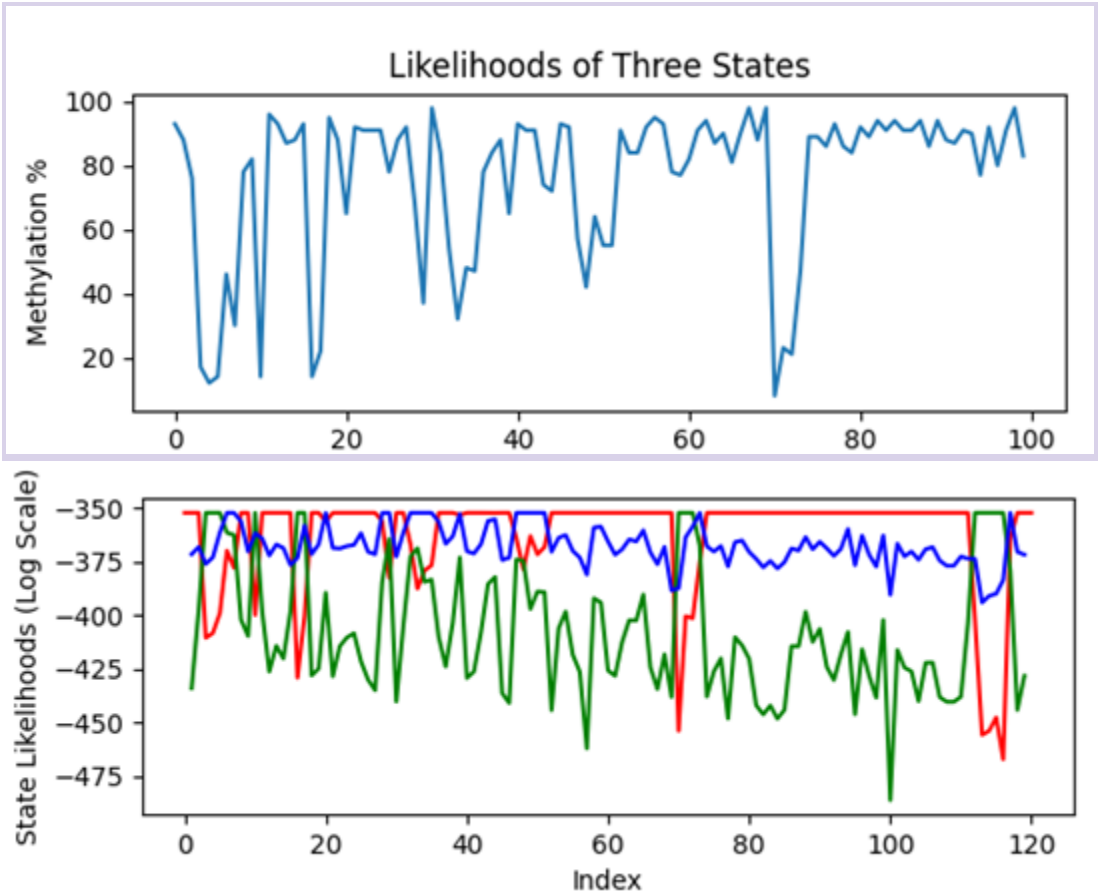
During the Data analysis and decomposition we’ve noticed there are many short bursts of one to three positions which were unmethylated, but all their surrounding was methylated. We decided to take a closer look on the phenomenon and try to include it in our model.

```
Prostate-Epithelial-Z0000045G   : 9999989999100000000000001000000022999989999999999999999
Prostate-Smooth-Muscle-Z0000041Y: 9998999995859216999999799499599998999999999999899899899
Blood-B-Mem-Z0000041J           : 9997989999699786798997498499898999999999999999999999999999
Blood-B-Mem-Z0000041K           : 9996989989695889989885794999998999999999999999999888999
Blood-B-Z000000TX               : 9996999999898799999999799999899999989989999799999998999
Blood-B-Z000000UB               : 9998999999998889999999899999999999999999999999999999999999
```

For this purpose we introduced a new state, “Other”, and calculated it’s transition parameters using the EM algorithm and fine tuning to receive the following outcomes



As represented in the automouta the Other state is not very common nor very stable, were the probability to persist in this state is only 0.15% and it very soon moves forward almost exclusively to the High state. We were intrigued and decided to check if the phenomenon is real or just a fluke in the data, we plotted the likelihood of the three states on course of 120 CpG bp and observed that the Other state explains well all the areas were both the High nor the Low states explains perfectly, and overall the additional state improve the prediction accuracy of the model.



Top: Methylation percentage of CpG sites in a representative area
Bottom: log-likelihood of the High (red), Low (green), and Other (blue) states at the same indexes under the trained model

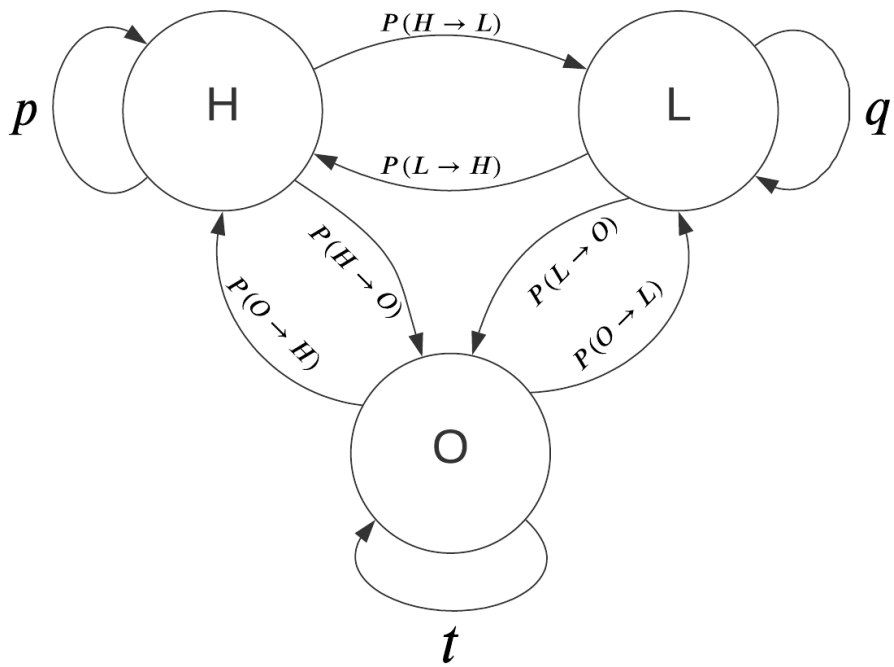
We derived a few hypothesis that included that these areas are Promoters for genes adjacent to them or even places for regulation of gene expression that can be up to thousands of bp away of them, another hypothesis is these short bursts are sample noise, and the enchanted likelihood comes from overfitting of the model. All three possibilities will have to be further reviewed in future study with more data.

Discussion

Forcing zero probability for transition between low-other

We can characterize the data in one more feature. Looking at some random places in the data it seems like the low areas tend to be continuous, and in addition we have some one-length jumps from high areas to low. So it could be better to model the data that way. In order to do so all we need is at initialization to give zero probability to transition from LOW state to the third state, and

vise versa. We will get that at each cycle the model will give zero probability for such a transition, and it will remain zero at the end.



What did our method accomplish?

Our HMM seems to predict reasonably well when a region is mostly methylated and when it is mostly unmethylated. This can be used to determine markers. For example we could create annotation files for each kind of tissue to see when it is mostly methylated and what regions it is mostly unmethylated. Then we could see where there are disagreements. For example if we find a region that is mostly unmethylated only in prostate samples, then this region could be a marker for prostate.

Abstract; Proofs

Maximum Likelihood Estimation of a binomial Distribution

If we model methylation as a binomial distribution then we can call p the probability of a read at a particular site to be methylated or not. Let C_l be the number of methylated observations in sample l and n_l be the total observations. So the likelihood is:

$$\prod_l \binom{n_l}{C_l} p_l^C (1-p)^{n_l - c_l}$$

and the log-likelihood is:

$$\sum_l \log \binom{n_l}{C_l} + C_l \log(p) + (n_l - C_l)\log(1-p)$$

Taking the derivative and setting equal to zero we get:

$$\sum_l \frac{C_l}{p} - \frac{n_l - C_l}{1-p} = 0 \Rightarrow \sum_l C_l(1-p) - p(n_l - C_l) = 0$$
$$\Rightarrow \sum_l C_l - C_l p - p n_l + p C_l = 0 \Rightarrow \sum_l C_l - n_l p = 0$$
$$\Rightarrow p = \sum_l \frac{C_l}{n_l}$$

Calculating the sufficient statistics

Let k be the hidden state. Each hidden state has it's own binomial destitution parameters.
Let C_k/ T_k be the number of methylated/unmethylated observations of all sites which have hidden state k. We have shown that the maximum likelihood estimator of the parameters of the binomial distribution of hidden state k are

$$p = \frac{C_k}{C_k + T_k}$$

Thus the sufficient statistics for finding the ML estimator are C_k and T_k. Since this is unlabelled data and we do not know which sites belong to which hidden states, we can estimate $\mathbf{E}[C_k]$ and $\mathbf{E}[T_k]$.
Let x^l be the set of sequences from which we want to learn our HMM parameters (i.e. each x^l is a WGBS sample).
Then x^{l_j} refers to read j in sample l
$x_i^{l_j}$ refers to read j in sample l at CpG site i.
Thus:

$$C_k = \sum_{l=1}^{M} \sum_{j=1}^{R} \sum_{i=1}^{N} \mathbb{1}\{S_i = k, x_i^{l_j} = C\}$$

and so:

$$\mathbf{E}[C_k] = \sum_{l=1}^{M} \sum_{j=1}^{R} \sum_{i=1}^{N} P(S_i = k, x_i^{l_j} = C | x^{l_j})$$

Where the conditional is provided since x is observed.

$$\begin{aligned}
\mathbf{E}[C_k] &= \sum_{l=1}^{M} \sum_{j=1}^{R} \sum_{i=1}^{N} P(x_i^{l_j} = C | x^{l_j}, S_i = k) * P(S_i = k | x^{l_j}) \\
&= \sum_{l=1}^{M} \sum_{j=1}^{R} \sum_{i=1}^{N} \mathbb{1}\{x_i^{l_j} = C\} * \frac{F_k(i) B_k(i)}{P(x^{l_j})} \\
&= \sum_{l=1}^{M} \sum_{i=1}^{N} C_i^l \frac{F_k(i) B_k(i)}{P(x^{l_j})}
\end{aligned}$$

Where C_i^l is the number of methylated observations in sample l at site i and $F_k(i)$ and $B_k(i)$ are the forwards and backwards tables from the Baum-Welch algorithm learned in class and $p(x)$ is the likelihood of the sample.