CBIO Project

Yotam Constantini, Ela Fallik, Guy Segal

February 2021

Introduction

Methylation: One form of RNA expression regulation is through methylation of the 5th position of cytosine in the DNA. In mammals, this change is made almost exclusively in cytosine-guanine nucleotide (CpG) pairs. CpG methylation has been shown to play an important role in many parts of normal development [1].
Out of about 28 million CpG sites in the human genome, 60-80% are methylated. However, in CpG islands, which are short regions of the genome with high CpG concentration (2-3% vs <1% of nucleotide bipairs), the methylation percentage is typically very low. Those islands routinely appear near promoters and enhancers.
Methylation is generally static in an individual, although there are some minor changes across cell-types, throughout life, and in mitosis [2]. More drastic changes in methylation patterns have been linked to many diseases and disorders [3].

Co-methylation and Distances: Since the importance of methylation for normal development was established, the mechanisms that determines methylation throughout the genome and during mitosis was under intensive research. Co-occurrence of methylation across the genome, such as seen in CpG islands, indicates a non-stochastic process, and may help to understand patterns of methylation. Several studies have found a high tendency for co-methylation in proximate CpG sites, and a decrease in co-methylation as the distance between sites increases [4].

The Project Goal: Our goal is to investigate the phenomenon of methylation co-occurrence in our data. Specifically, we aim to study the effect of distance on the interactions between methylated and unmethylated neighboring CpG sites, and how those interactions change across different cell types and in different parts of the genome (Specifically, different chromosomes and inside CpG islands). Our work can be divided into two major parts: statistical analysis and predictive model. In the **statistical analysis** part, we ask questions such as: Can we see a decrease in co-methylation as distance increase? If so, is this decrease exponential? Can we see a difference between samples and genome sections? In the second part we perform **predictive modeling** of the interaction between CpG sites, using a continuous time Markov chain (CTMC). This model assumes an exponential decrease in co-methylation as the distance between CpG sites increase, and attempts to predict the co-methylation patterns in different parts of the genome.

Methods

Data

We worked with two different cell types - Heart-Cardiomyocyte and Liver-Hepatocytes, three samples each.
Wet procedure: Tissue was dissociated, and cells sorted using FACS, DNA extracted and went through bisulfite treatment. After fragmentation a library was prepared from the DNA and went through paired ends (2×150) next-generation sequencing using Illumina.
Computational processing: Reads were mapped to hg19, filtered for low quality and mapping score, and 3 bases were trimmed from either sides. CpG genomic sites that appeared as CG and TG were marked as methylated (C) and unmethylated (T) respectively. Paired reads were joined, overlap-disagreements and CpG sites in unsequenced gap marked (.). To save space, reads were saved in g-zipped ".pat" format: Using a CpG sites indexed, reads are translated to the indexed methylation pattern and aggregated by similar patterns and start site:

Col1	Col2	Col3	Col4
chromosome	CpG index of first site	methylation pattern /[CT.]*/	number of copies

Additional files: A dictionary file was used to map CpG indices to locations in the genome, and to extract true distances. A CpG islands dictionary was used to identify CpG sites inside/ outside CpG islands.

Empirical Model

Given the data, we were interested to see if there is a pattern in the correlation of methylation between adjacent CpG sites. We summed the empirical occurrence of transitions *methylated* $\rightarrow$ *methylated* ($M \rightarrow M$) and *unmethylated* $\rightarrow$ *unmethylated* $U \rightarrow U$ between **adjacent** CpG sites, for different distances. We denote by $\hat{P}_n[M \rightarrow M]$ the empirical proportion of transitions from methylated to methylated sites of distance n, and similarly for $U \rightarrow U, M \rightarrow U, U \rightarrow M$. We regard the proportion as the transition probability, and denote the matrix for some distance n by

$$\hat{P}(n) = \begin{bmatrix} \hat{P}_n[M \rightarrow M] & \hat{P}_n[M \rightarrow U] \\ \hat{P}_n[U \rightarrow M] & \hat{P}_n[U \rightarrow U] \end{bmatrix}$$

Transition Matrices Extraction

First, an index matrix was created from each pat file, denoting the counts of each transition type for each pair of neighboring sites in the sample. The distance between each pair was extracted from the CpG indices dictionary. Throughput this work, we only considered adjacent sites with distance $n \leq 200$.
Next, for each group of CpG sites (whole genome, split by chromosomes, inside/outside CpG islands) we extracted the counts for the appropriate pairs from the index matrix, summed them by distances n, and evaluated the probability matrix $\hat{P}(n)$ for each n with more then 100 counts total.

Heuristic Exponential Fit

In order to demonstrate an exponential decay in the data, and examine the differences between genome segments, we heuristically fitted an exponential function to the empirical data:
For a data set $\forall_n, p_n = \hat{P}_n[i \to j]$ (for some $i, j \in [M, U]$) we fit an exponential function

$$\begin{aligned} f(n) &= N_0 \cdot \exp(-\lambda \cdot (n-2)) + C \\ \text{s.t. } & f(0) = \max p_n \\ & f(52) = p_{50} \\ & f(\max n) = \min p_n \end{aligned}$$

Note that the expression $(n-2)$ comes from the fact that the minimal distance between two pairs is 2. This is an entirely heuristic fit, as there's no reason to prefer $n = 52$ here. From those conditions we get the following constants:

- $N_0 = \max p_n - C$
- $\lambda = -\frac{1}{50} \ln \frac{p_{50} - C}{N_0}$
- $C = \min p_n$

We fitted two function with this process, one to the data set $\forall_n, \hat{P}_n[M \to M]$ and the other to $\forall_n, \hat{P}_n[U \to U]$.

Log Likelihood as a Score Function

Denote by $\hat{C}(n) \in \mathbb{R}^4$ the row vector of empirical counts of transition $M \to M, M \to U, U \to M, U \to U$ for distance n. Given a count matrix $\hat{C} \in k \times 4$ and probability matrix $\hat{P} \in k \times 4$ for k different distances, we score the compatibility of the probability matrix to the observed counts data using the log-likelihood function:
We first notice that the probability (given this probability matrix) to see the observed counts in a distance n, for a certain transition type (say, $M \to M$), is

$$\mathbb{P}(\hat{C}_n[M \to M] | \hat{P}_n[M \to M]) = (\hat{P}_n[M \to M])^{\hat{C}_n[M \to M]}$$

and the mutual probability of the observed counts in a distance n for all transition types, denoted $\mathbb{P}(\hat{C}_n | \hat{P}_n)$, is therefore the product of those probabilities.
Now, for a sequence of distances $n_1, ..., n_k$, the mutual probability of the observed counts for all distances, denoted $\mathbb{P}(\hat{C} | \hat{P})$, is the product of all the probabilities $\forall_i, \mathbb{P}(\hat{C}_{n_i} | \hat{P}_{n_i})$. Taking a log, we get the log-likelihood function

$$\text{LL}(\hat{P}; \hat{C}) = \sum_{\substack{1 \le i \le k \\ 1 \le j \le 4}} \hat{C}_{n_i}[j] \cdot \log \hat{P}_{n_i}[j]$$

In order for the scores achieved from different cell types and genome segments to be comparable, we normalize the log likelihood by the total amount of counts for the segment (sum of all entries of $\hat{C}$). We denote this score by normalized-LL$(\hat{P}; \hat{C})$.

Predictive Model

We modeled the methylation interaction between CpG sites using a CTMC, assuming each site is affected only by its adjacent sites. Formally, there exists $\alpha, \beta \in \mathbb{R}_{<0}$ for which

$$\frac{\partial P(n)}{\partial n} = P(n) \cdot R$$

where

$$R = \begin{bmatrix} \alpha & -\alpha \\ -\beta & \beta \end{bmatrix}$$

As we learned in class, this formulation gives

$$P(n) = e^{n \cdot R}$$

Learning

Using the normalized log-likelihood as a measure of the performance of our model we would like to obtain a rate matrix, $\hat{R}$, such that

$$\hat{R} = \underset{R}{\text{argmax}} \ \text{normalized-LL}\left(\left[e^{n_1 \cdot R}, ..., e^{n_k \cdot R}\right]; \hat{C}\right)$$

for the convenience of using **gradient descent**, we used the negative-normalized log-likelihood. That is, we would like

$$\hat{R} = \underset{R}{\text{argmin}} \ -\text{normalized-LL}\left(\left[e^{n_1 \cdot R}, ..., e^{n_k \cdot R}\right]; \hat{C}\right)$$

In order to solve this optimization problem we used the the gradient descent algorithm - for $\theta \in \{\alpha, \beta\}$, and $\gamma \in \mathbb{R}$ (learning rate), we initialize θ_0 randomly and update

$$\theta_{t+1} \leftarrow \theta_t - \gamma \cdot \frac{\partial \, \text{negative-normalized-LL}}{\partial \theta}$$

We used the Adam algorithm variation of gradient descent [5], that improves the optimization process by estimating the momentum of the dynamics. We omit the full formulation of Adam for brevity. Additionally, we used the automatic differentiation package JAX [6] in order to perform gradient descent.

Shrinkage and Expansion

Given a model with probabilities per distance, we can now ask for any neighboring pair of CpGs (S_j), what is the likelihood of seeing the empirical transitions assuming different distances:

$$L_i(S_j) \ = \ P\left(\textit{empirical transitions}(S_j) \mid \textit{distance} = i\right)$$

Calculating $L_i(S_j)$ for all CpG pairs according to all of the distances, we can ask, for each pair, what distance parameter maximizes the likelihood ($D_{ML}(S_j) = \text{argmax}_i(L_i(S_j))$) for that pair, and compare it to the true distance ($D_T(S_j)$) between the CpGs. The log ratio $LDR(S_j) = \log\left(\frac{D_{ML}(S_j)}{D_T(S_j)}\right)$ indicates the behavior of the specific CpG pair compared to the overall behavior, and can be viewed as shrinkage or expansion of the distance. To ensure stability of the results, we looked at pairs with coverage of fifteen or higher.

Results

Empirical Analysis

For each pair of adjacent CpG sites in each of our samples, we counted the number of observed transitions from each type (MM, MU, UM, UU). Next, we normalized those counts to get a probability matrix for each pair. We discarded pairs for which no transition originating from M or U was observed. From this process, we obtained matrices that give each pair of adjacent CpG sites a probability for each type of transition.

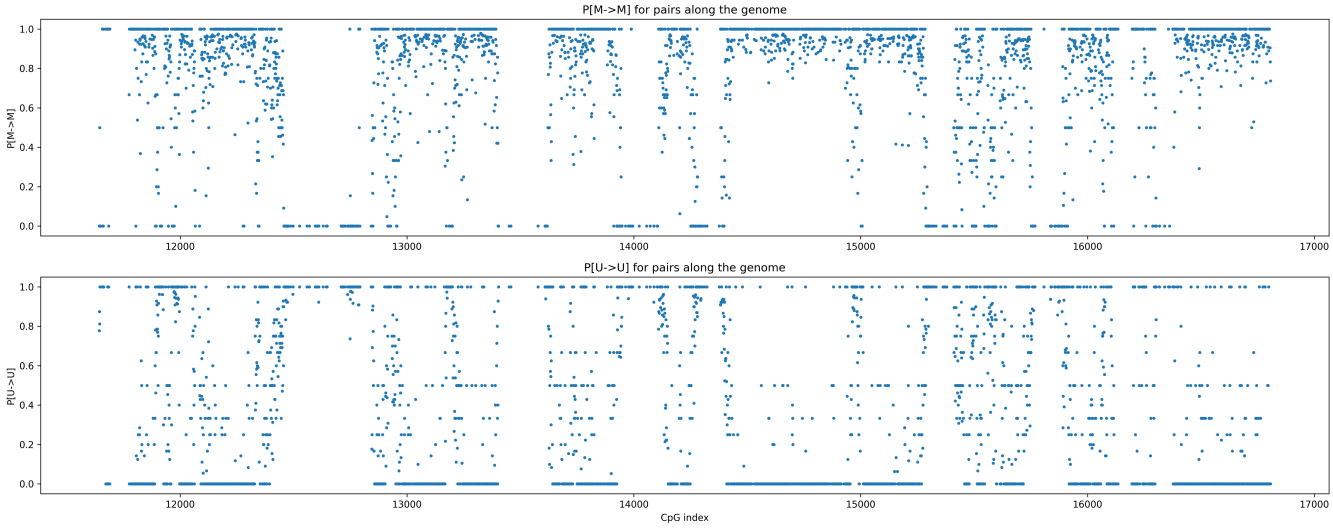

Figure 1: Empirical transition probabilities $P[M \to M]$ and $P[U \to U]$ for adjacent CpG sites in a segment of the genome, estimated from sample Heart-Cardiomyocyte-Z0000044G

For many pairs we observed only one type of transition from a state, indicating cell-type consistency. We can see differences in the transition probabilities between sections of the genome (Figure 1). For example, we see sections in which most pairs received a high $U \to U$ probabilities, compared to a low $M \to M$ probabilities. This suggests that most indexes in those sections are unmethylated, indicating CpG islands.

Next, we summed the counts of each types of transitions according to the distances between the CpG pairs, and normalized to get a probability matrix $\hat{P}(n)$ for each distance n. We performed this calculation for all pairs across the genome, inside CpG islands, outside CpG islands, and in each chromosome separately. In order to reduce noise levels, we discarded (for this section only) distances for which less than 100 counts were observed. As CpG islands are characterized by densely positioned CpG sites, there was only enough data to reach $n = 100$.

Additionally, in order to visually demonstrate an exponential decay, we fitted an exponential function of the form $f(n) = N_0 \cdot \exp(-\lambda \cdot (n-2)) + C$ to the data (Methods). The decay factors calculated for the data presented in figure 2 are detailed in table 1.

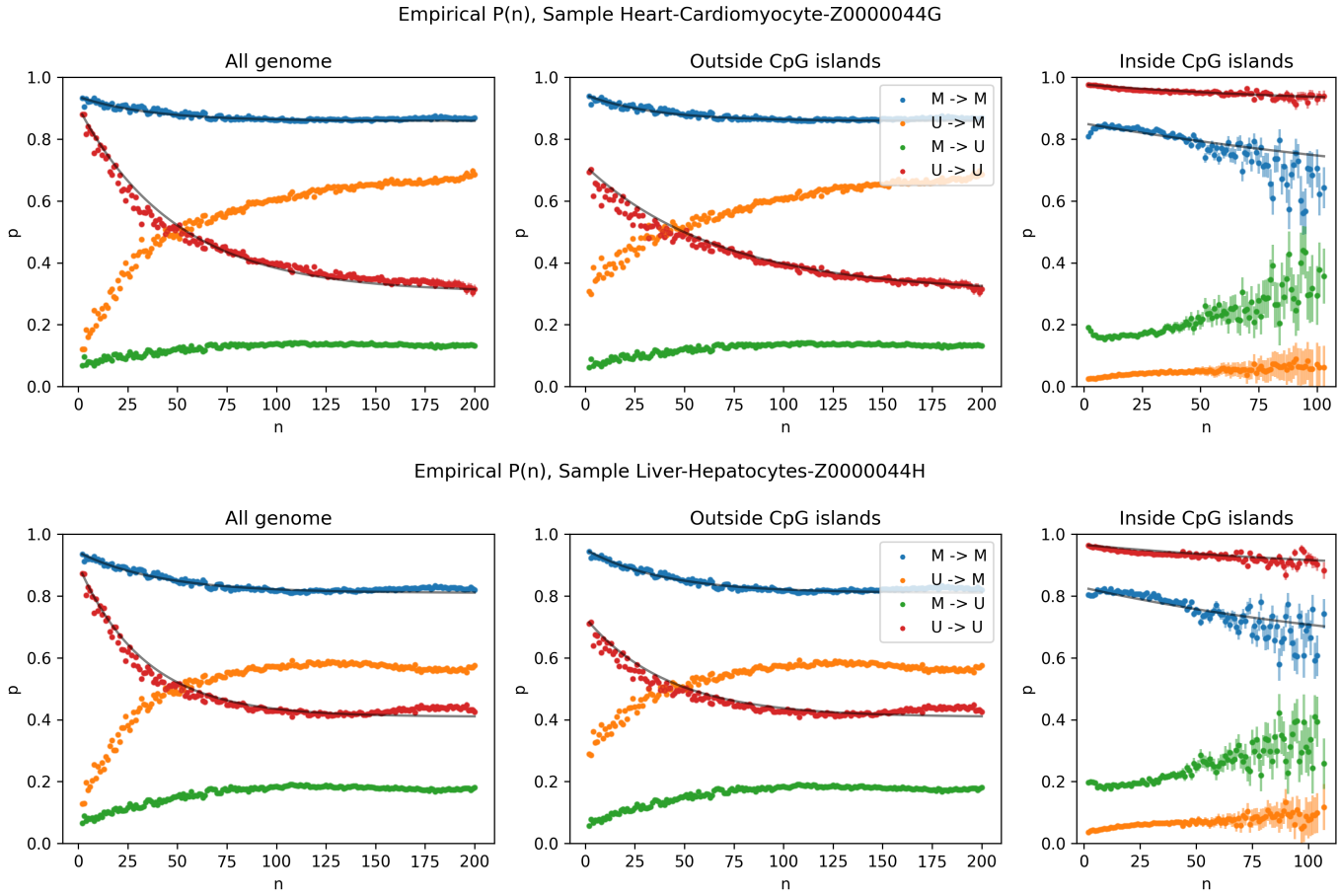

Figure 2: Empirical transition probabilities, estimated from different genome segments of two samples from different cell types. Confidence intervals were calculated using Z-score. An heuristic exponential fit is plotted on top of the probabilities $P[M \to M]$ and $P[U \to U]$ (black).

		All Genome	Outside Islands	Inside Islands
Heart-Cardiomyocyte Z0000044G	$M \to M$	0.027	0.028	0.004
	$U \to U$	0.020	0.014	0.009
Liver-Hepatocytes Z0000044H	$M \to M$	0.024	0.025	0.006
	$U \to U$	0.029	0.024	0.006

Table 1: Decay factors calculated for the data presented in figure 2.

Looking at the probabilities evaluated across the genome (Figure 2, left), we see an exponential decay in $M \to M$ and $U \to U$ probabilities as distance increases. For the Heart cells, there is a smaller decay factor for the $U \to U$ transition, suggesting that the probability for a transition from a methylated site is less distance-dependant than from an unmethylated site.
We can see small differences between cell types in large distances. Specifically, in most liver cells there are higher $U \to U$ probabilities compared to heart cells. The decay factor in the liver samples is bigger for $U \to U$ while smaller for $M \to M$ compared to heart cells. However, the exponential fit is less accurate for long distances (150-200). Across samples from the same cell type, the results were almost identical, except for one

liver cell sample (Z0000043Q) which seems to be suspiciously similar to the heart samples.

There are small differences at short distances between the probabilities evaluated across the entire genome and those evaluated outside CpG islands (Figure 2, center). This is due to the fact that most of the pairs within CpG islands are located at small distances, so their exclusion affects almost exclusively those distances. The $U \to U$ probabilities are 20% lower (overall) after removal of data from CpG islands, with a smaller decay factor. Those differences were seen in all samples.

Within CpG Islands, however, the situation is vastly different compared to outside those segments (Figure 2, right). The most significant difference is found in the $U \to U$ probabilities, which are close to 1 inside the islands, compared to 0.8-0.4 outside islands (depending on distance), and have a very low decay factor (0.009 / 0.006). The $M \to M$ probabilities are lower than in the rest of the genome, and do not fit to an exponential decay plot.
As noted, there was little data for large distances within CpG islands, and even in relatively short distances (50-100) the probabilities noise levels are significant. We notice that there is greater noise in the probabilities of transitions originated from methylated sites compared to unmethylated, indicating a lower number of methylated sites overall (as characteristic of CpG islands).
Overall, most pairs inside the islands are located at short distances, have a higher probability of transitioning to an unmethylated site, and a probability close to 1 of staying in this state. This fits with what we know about CpG islands - the density of CpG sites is high, and most sites are unmethylated. There were no notable differences between cell types, samples and chromosomes (up to chromosome size effects).

We conclude that for $M \to M$ and $U \to U$ probabilities estimated across the genome and outside CpG islands, exponential decay can be seen as the distance between adjacent CpG sites increases. Therefore, a predictive model that assumes exponential decay such as CTMC is likely to fit this data well. Regarding CpG islands, where some of the estimated probabilities do not meet the exponential decay assumption and are very noisy, we don't expect to see a good fit with a CTMC model.

Predictive Model

Synthetic Data

In order to validate our method we first apply it to synthetic data. That is, we fix some α, β as decay rate, generate probabilities $P(n)$ for $1 \leq n \leq 200$ and observe whether the optimization method can retrieve those parameters (Figure 3). Inspecting the progress of the gradient decent, we see α, β convergence after 80 iterations (Figure 4). We conclude that for data that fits an exponential decay model, and in the absence of noise our algorithm performs well.

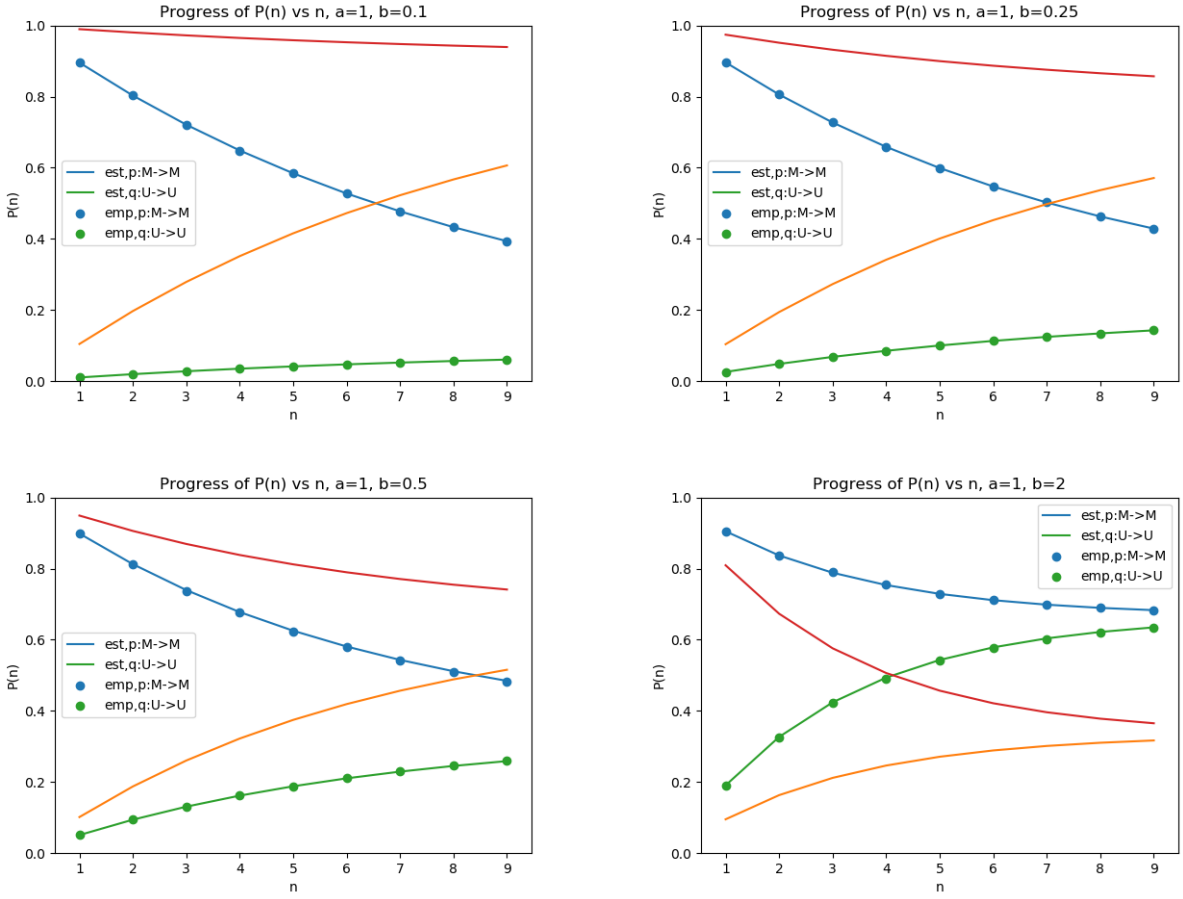

Figure 3: Reconstruction of $P(n)$ from synthetic data. Dots are generated $P(n)$ from the synthetic data and lines are the predicted values from our optimization algorithm. Red and orange lines are complementary $U \to M$ and $M \to U$.

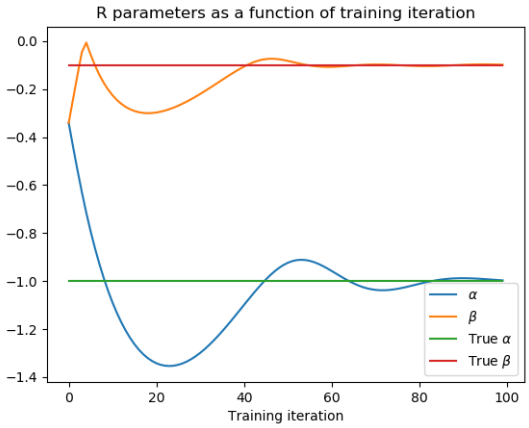

Figure 4: Training dynamics of α and β

Real Data

Establishing that our approach works for synthetic data, we turn to apply it to our real data. We ran the fitting algorithm for each cell type, split by chromosomes that were

further split by inside/outside CpG islands, and plotted the resulting α and β parameters. Two phenomena can be readily observed - parameters learned for inside islands cluster differently from those of outside islands and whole chromosome, and the overall likelihood is higher for the islands.

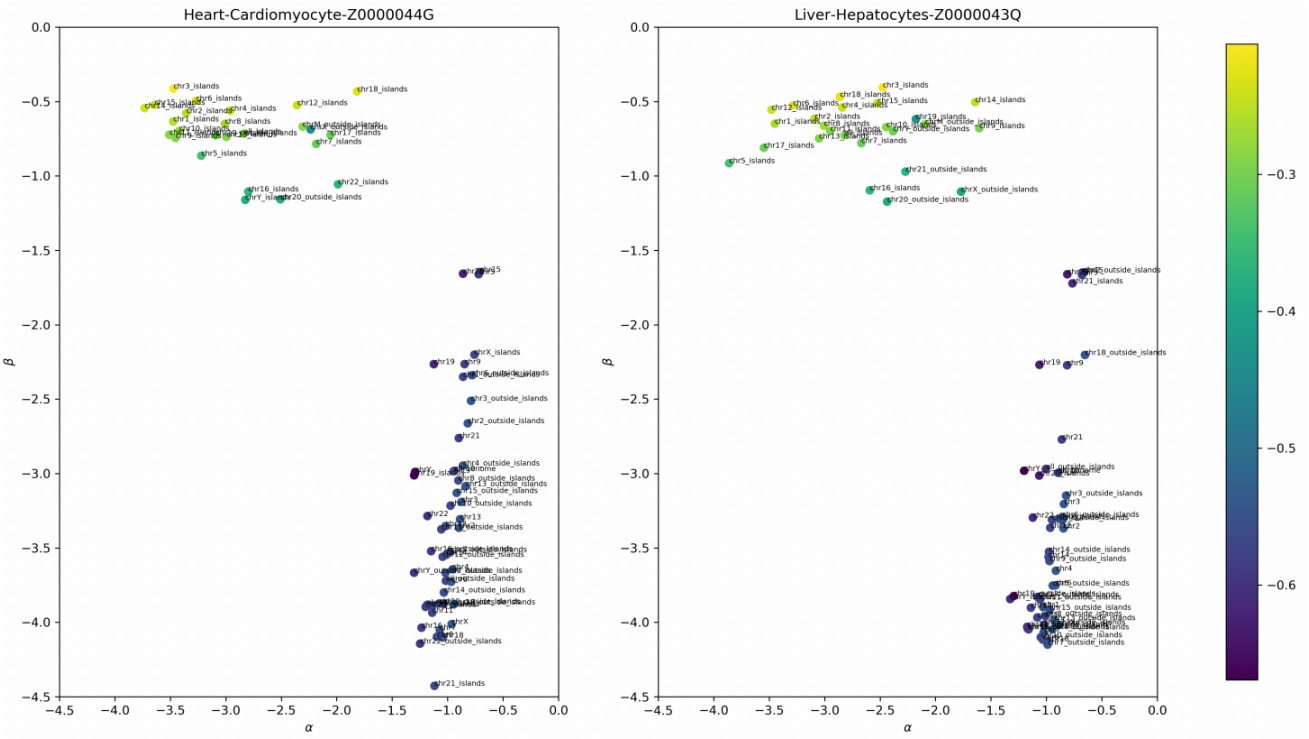

Figure 5: Learned parameters for different segments from two cells - a heart cell and a liver cell. Color of dots is log-likelihood achieved in training.

Both phenomenon can be explained by the fact that inside the islands most transitions are $U \to U$ and concentrate on smaller distances. This yields α, β that are obviously different from the ones learned from segments outside the islands but also results in a function that is easier to approximate where it matters most - the smaller distances. Since inside the islands the function is easier to approximate the algorithm achieves a better score.

Using the log-likelihood as a score function drives the algorithm to focus on the approximation for distances where there are a lot of transitions. This fact can be seen in the following plots -

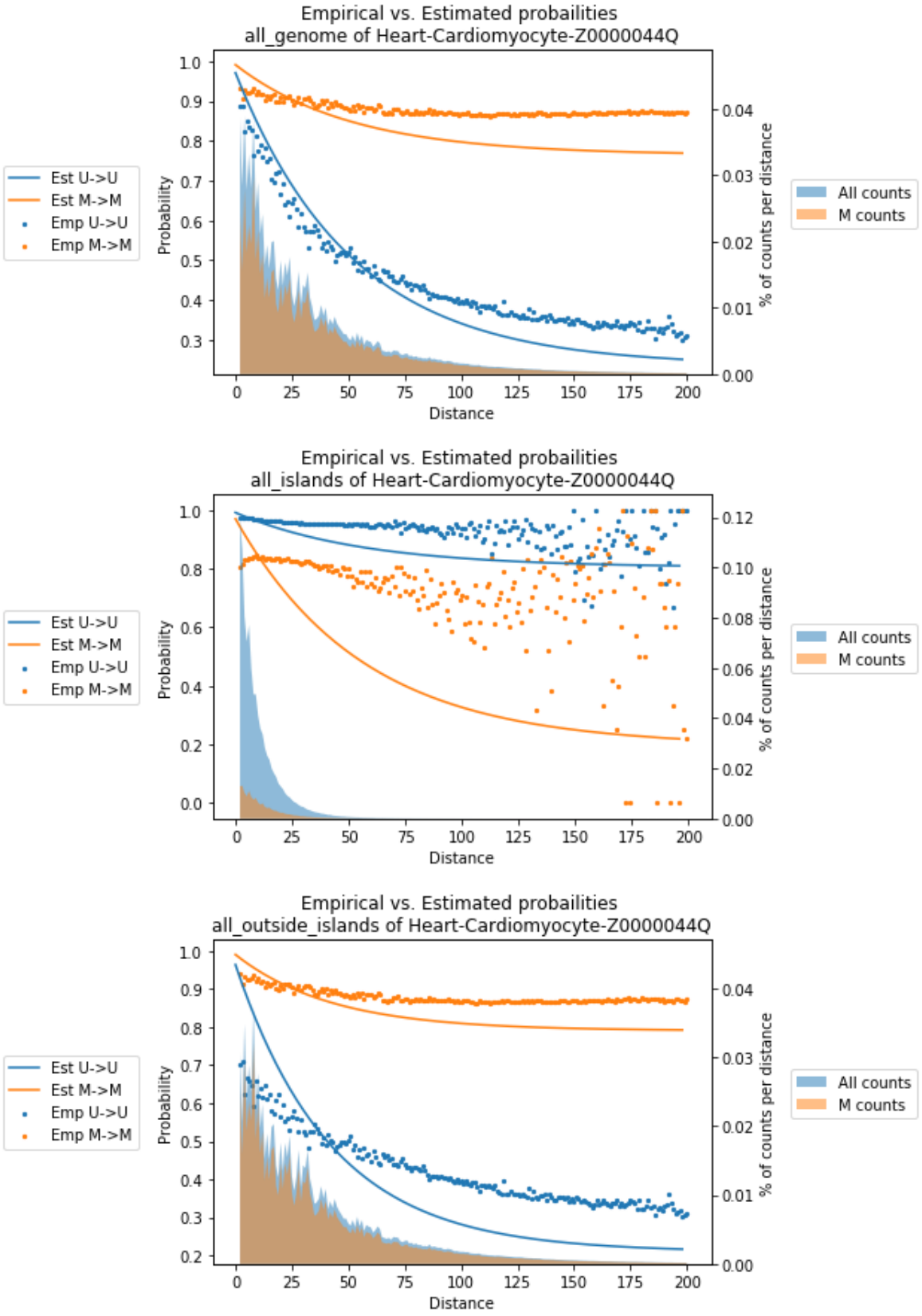

Figure 6: Estimated transition probabilities vs. empirical ones (left y-scale) and percentage from total counts per distance (right y-scale) for heart cells. Blue is percentage of all transitions per distance and orange is percentage of only transition that originated in a methylated site. The difference between the areas is the percentage of transitions that originated in unmethylated sites.

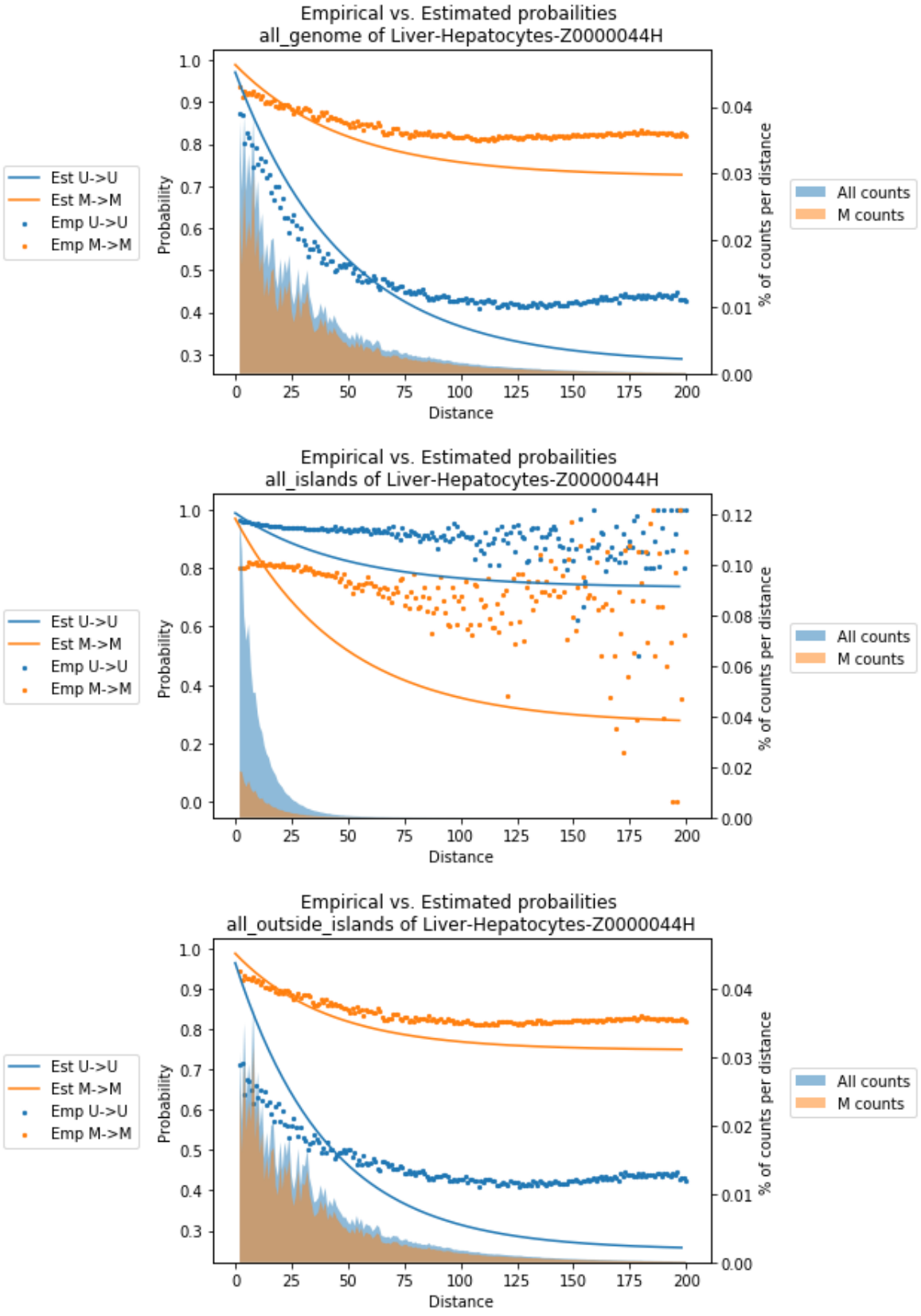

Figure 7: Estimated transition probabilities vs. empirical ones (left y-scale) and percentage from total counts per distance (right y-scale) for liver cells. Blue is percentage of all transitions per distance and orange is percentage of only transition that originated in a methylated site. The difference between the areas is the percentage of transitions that originated in unmethylated sites.

In both of the figures above the left y-scale is the probability (estimated and empirical) and the right y-scale corresponds to the percentage of counts per distance. We can see quite clearly that for the islands segments transitions from methylated sites are very rare and thus the algorithm puts more weight on the approximation of the $U \to U$ part.

Next we set out to examine, under our CTMC model, how well do parameters learned on one segment of the genome fit the others.

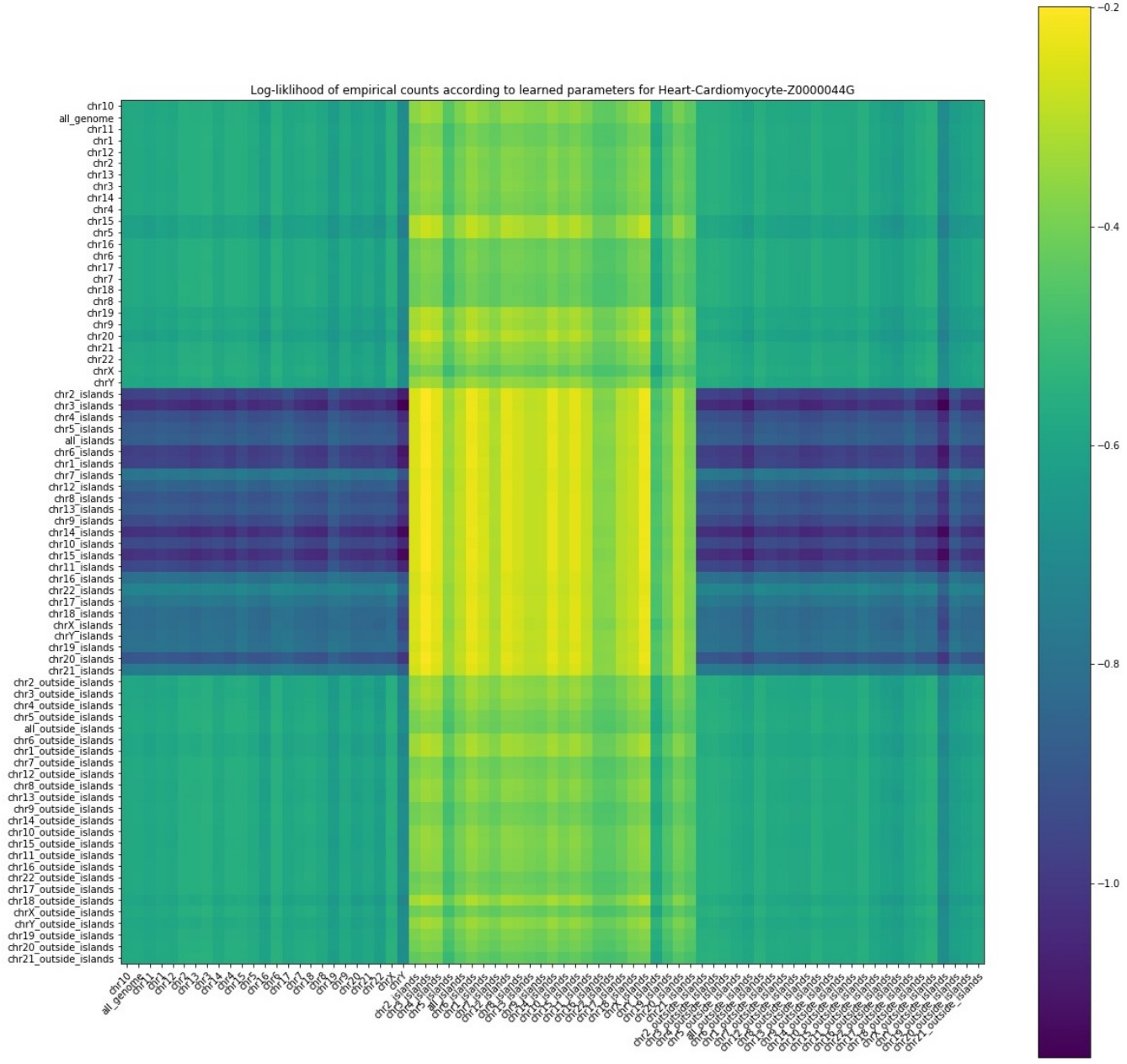

Figure 8: Log likelihood for empirical count matrices, calculated based on parameters learned on different genome segments.

The first observation is that the CpG islands cluster separately from segments of entire chromosomes and segments outside islands. Training log likelihood of parameters learned on the CpG islands outperform the rest. Despite that, the test log likelihood

(their performance on segments outside islands) fail to match the ability of parameters learned outside islands. Interestingly, the test log likelihood of entire chromosomes / outside islands is higher than their training log likelihood.

Shrinkage and Expansion

We calculated the likelihood of the observed counts for different CpGs in the genome according to empirical transitions probabilities.

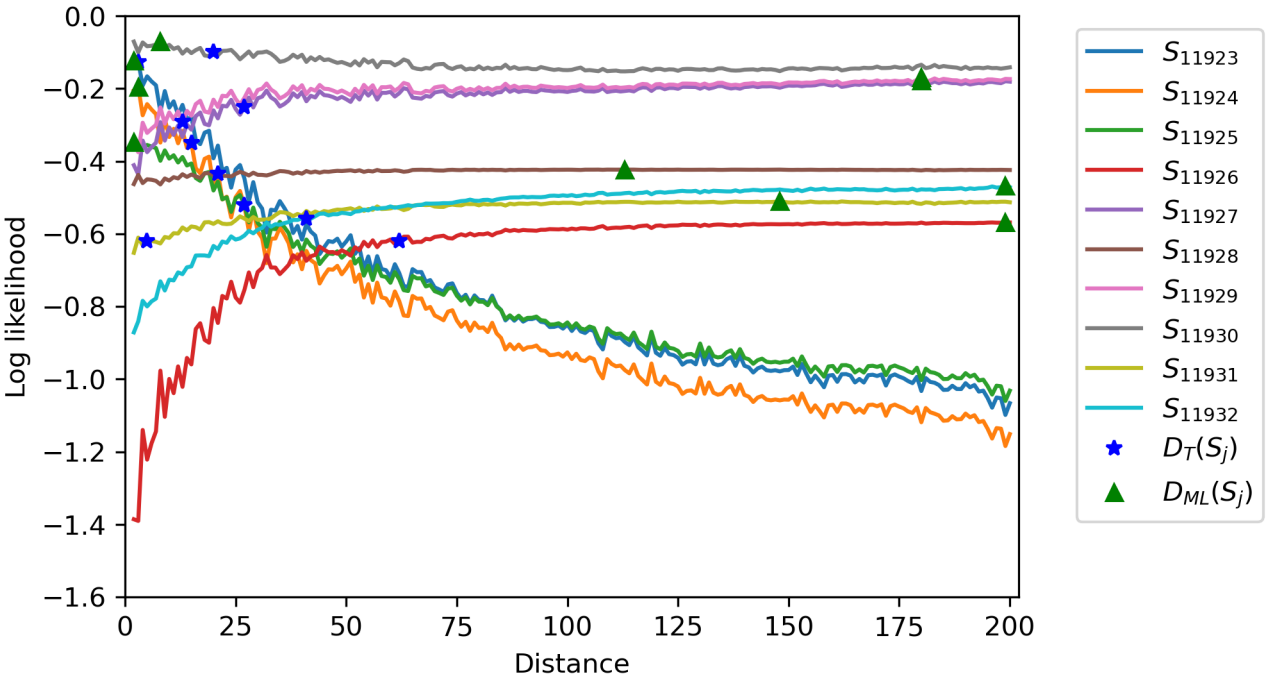

Figure 9: normalized-LL$(\hat{P};\hat{C})$ of observed counts according to empirical transitions

We then moved on to examine the relationship between the true distances and the obtained ML distances throughout the genome. We made a histogram of the accuracy, measured as $|D_{ML} - D_T|$ blue dots), and compared it to the difference between the true distances and a distance sampled randomly $2 \leq d \leq 200$ (orange line). To control for effects caused by the skewed nature of the true distances, we shuffled the distances and calculated $|shuffled(D_T) - D_T|$ (green line). The three triangles below the graph indicate the mean of the absolute distance. We plot the spiky distribution of the empirical ML distances (91% concentrated on the red asterisks).

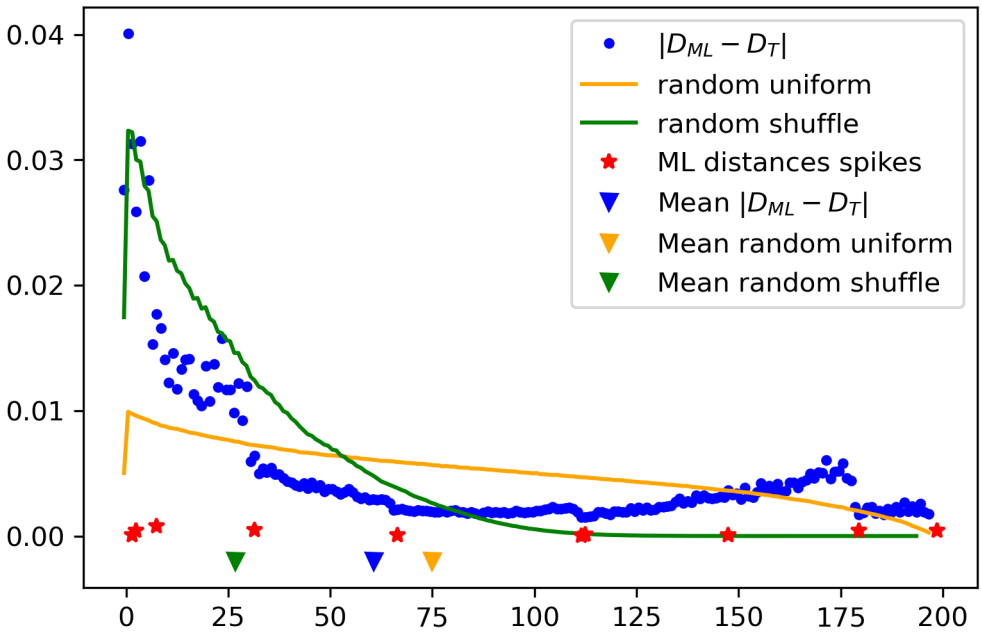

Figure 10: Accuracy of the ML-distance

The Global accuracy of the ML distance being shaky, we looked into specific regions of the genome (CpG islands marked in red asterisks).

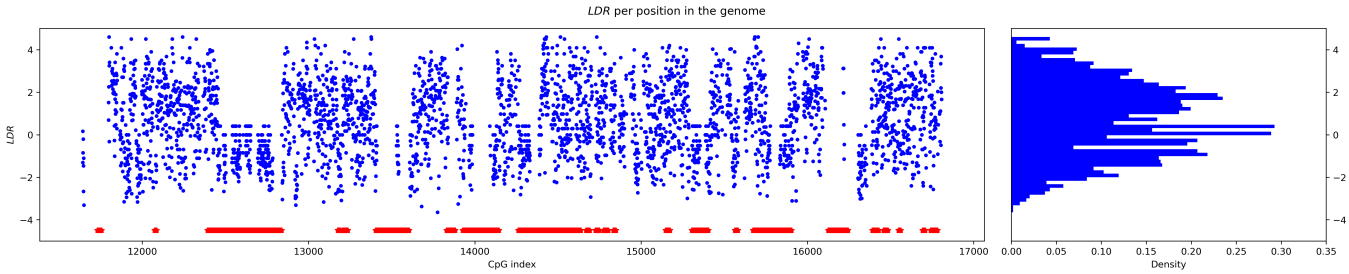

Figure 11: Likelihood distance ratio according to empirical transitions

As noise levels had a great effect on the prediction, we repeated the above steps, this time using the CTMC model parameters.

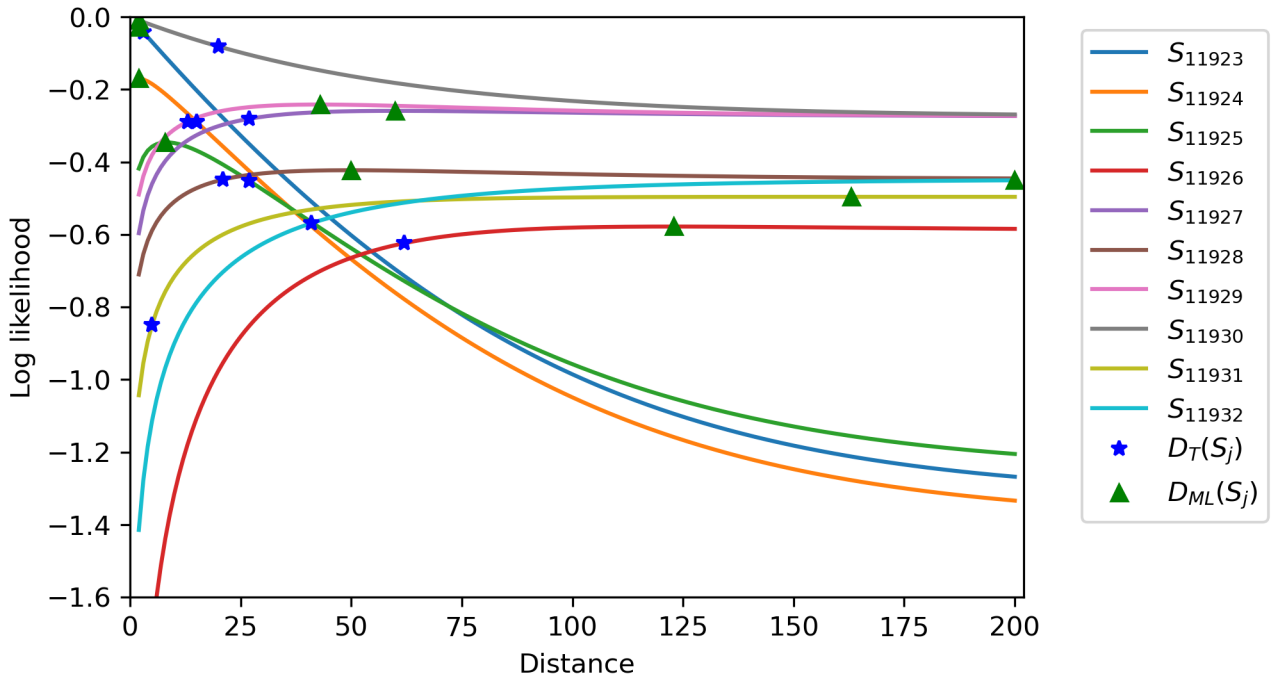

Figure 12: normalized-LL$(P;\hat{C})$ of observed counts according to CTMC based transitions

Accuracy histogram as before, we measured as $|D_{ML}-D_T|$ blue dots), and compared it to the difference between the true distances and a distance sampled uniformly $2 \leq d \leq 200$ (orange line). We again shuffled the true distances and plotted $|shuffled(D_T) - D_T|$ (green line). The three triangles below the graph indicate the mean of the absolute distance. Using the CTMC based transitions the ML distribution became smoother (70% concentrated on the red asterisks), and a lot closer to the true-distances distribution (data not shown).

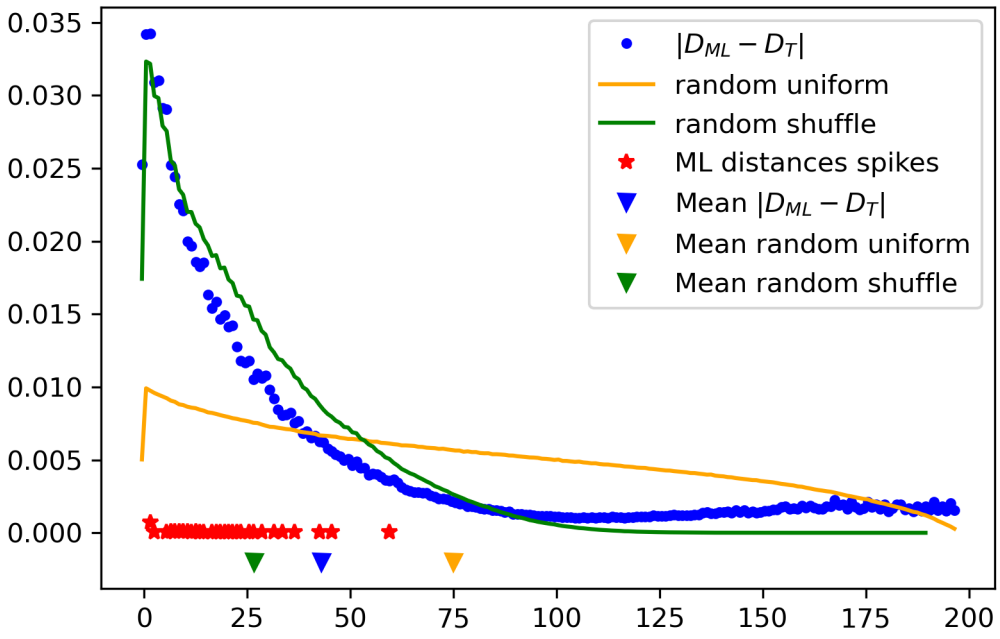

Figure 13: Accuracy of the ML-distance obtained by the CTMC based transitions

To complete the comparison of the model, we plotted the LDR according to the CTMC model based transition matrices.

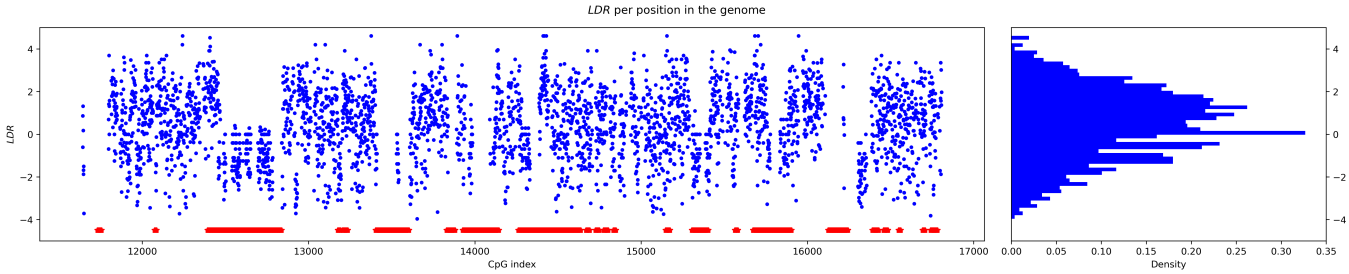

Figure 14: Likelihood distance ratio according to CTMC based transitions

Discussion

references

In this work we modeled co-methylation of adjacent CpG sites as a function of their distance. We've used two main types of models: empirical probabilities per distance, and a novel CTMC model. While both approaches showed improved prediction ability when utilizing the distance as a factor, the strongest effect we've observed on the prediction ability was that caused by CpG islands. CpG islands were good predictors for themselves, but had poor generalization ability for the rest of the genome. This supports the reported bi-modal nature of the methylation patterns. When looking outside CpG islands, distance played a major role in prediction ability.

Prediction

When examining the learned parameters from different segments of the genome two main clusters emerge - the islands cluster and the outside-island cluster. This corresponds the the analysis made in the empirical section of this work that concluded that those segments are very different from one another.

When observing the reconstruction of $P(n)$ from the CTMC model, one can notice the steady state proportion of the transitions, corresponding to long distances, is lower than the empirical. This phenomenon makes sense considering the normalized log-likelihood cost function - as the sites distances distribution is skewed to the left, the fitting of short distances has higher weight. When we re-ran the algorithm giving equal weights to the different distances, the steady state fitting was almost perfect (data not shown).

Shrinkage and Expansion

After modeling the effects of distance on co-methylation, the last part of our project was parameter estimation: given a set of neighboring sites, what distance gives the maximum likelihood of seeing their co-methylation profile? First we see that for the empirical modal, the noise caused concentration of ML distances in selected points in the range, concentration that smoothed when using the CTMC model. For both the empirical $\hat{P}(n)$ and the CTMC reconstructed $P(n)$ we see that CpG pairs with mainly $U \to U$ and $M \to M$ transitions get the maximum likelihood at the minimal distance of 2. On the opposite side of the graphs, many $U \to M$ and $M \to U$ transitions pulled the

ML distance - a rare event, but with significant impact on the tail of the $|D_{ML} - D_T|$ distribution (figure 10,13: distances 150 and above). To evaluate the use of ML distance, we first compared it to a uniform distribution of distances $2 \leq d \leq 200$. This comparison corresponds to a null hypothesis stating that the ML is indifferent to the true distances. The comparison showed significant added value of the CTMC $P(n)$ based ML distance, and a small added value when using the empirical $\hat{P}(n)$ based ML-distances (Figures 10, 13 - triangles). A slightly more challenging null hypothesis is that the distances are randomly sampled from the distribution of the true distances (Random shuffle - green line). Under this comparison, it seems both the CTMC $P(n)$ and empirical $\hat{P}(n)$ based ML-distances are hindering estimation efforts. In conclusion, the ML approach did not lead us far in trying to estimate the distances between adjacent CpG sites.

CpG islands

As CpG islands methylation is related to regulation, different islands throughout the genome are expected to be fully methylated or unmethylated, and this phenomena is hinted by the difference in LDR for different islands, as seen in figure 14. This raises a possibility for future work - splitting the CpG islands to methylated and unmethylated based on methylation percentage, and learning the parameters separately. Presumably, the fact that we measured co-methylation of adjacent sites helped us getting $U \rightarrow U$ parameters that were most likely solely based on unmethylated islands - thus probably identical to parameters we would obtain by estimating based only on unmethylated islands.

Project Code

https://github.cs.huji.ac.il/guyseg/CBIO-Project

References

[1] Z.D. Smith and A. Meissner. Dna methylation: roles in mammalian development. *Nat. Rev. Genet, 14 (3), pp. 204-220*, 2013.

[2] Dor Y and Cedar H. Principles of dna methylation and their implications for biology and medicine. *Lancet 392, 777–786*, 2018.

[3] K. Robertson. Dna methylation and human disease. *Nature Reviews Genetics 6, 597–610*, 2005.

[4] Affinito Ornella et al. Nucleotide distance influences co-methylation between nearby cpg sites. *Genomics 112.1: 144-150*, 2020.

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[6] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.