UNIVERSITY OF CALIFORNIA

Los Angeles

Modeling Site-Site Dependency in DNA Methylation Sequencing data

A thesis submitted in partial satisfaction
of the requirements for the degree
Masters of Science in Statistics

by

Wenbin Guo

2024

ABSTRACT OF THE THESIS

Modeling Site-Site Dependency in DNA Methylation Sequencing data

by

Wenbin Guo
Masters of Science in Statistics
University of California, Los Angeles, 2024
Professor Jingyi Li, Chair

DNA methylation is a crucial epigenetic modification on CpG sites, influencing gene expression and cellular function. Conventional analyses often neglect the intrinsic dependencies between adjacent CpG sites, limiting insights into underlying biological mechanisms and constraining their broader applicability. This thesis aims to model the site-site dependency in DNA methylation sequencing data using two complementary methodologies: a statistical approach using heterogeneous Hidden Markov Models (HMMs) and a machine learning approach employing Bidirectional Long Short-Term Memory (BiLSTM) networks.

The heterogeneous HMM extends the classical homogeneous HMM by incorporating genomic distance into transition probabilities, reflecting the biological intuition that adjacent CpG sites with closer proximity exhibit stronger dependencies. A parameter estimation procedure utilizing the Expectation-Maximization algorithm is derived to handle this extension. Simulation studies demonstrate that the heterogeneous HMM outperforms the homogeneous HMM in model fitting, parameter estimation accuracy, and capturing distance-related dependency patterns. When applied to whole-genome bisulfite sequencing (WGBS) data, the heterogeneous HMM provides a more accurate representation of methylation patterns, effectively capturing the diminishing dependency as genomic distance increases.

To address the limitations of HMMs in capturing complex and long-range dependencies, the thesis also introduces a deep-learning approach using BiLSTM networks. This model leverages the recurrent neural network architecture to implicitly learn sequential dependencies in both forward and backward directions. By incorporating a rich set of features—including methylation levels, genomic distances, and sequence context embeddings—the BiLSTM simultaneously captures marginal methylation probabilities and preserves site-site dependencies. Simulation studies and WGBS data analyses demonstrate its superiority over both homogeneous and heterogeneous HMMs in accurately aligning with marginal methylation levels and effectively preserving the intricate dependency patterns.

These two complementary approaches enhance the ability to characterize methylation pattern dynamics observed in real data. The heterogeneous HMM offers interpretability in modeling distance-dependent dependencies, while the BiLSTM provides flexibility to incorporate various features and capture complex dependency patterns. The thesis also outlines future directions to enhance these methodologies, including applying the frameworks to diverse datasets for broader generalizability, improving the computational scalability of the heterogeneous HMM for large-scale datasets, leveraging explainable machine learning techniques to identify key features driving methylation concordance, and exploring advanced generative models such as transformers and diffusion models for methylation pattern modeling. By effectively capturing site-site dependencies, these methods show promise for practical applications, such as imputing missing values in sparse datasets and improving the detection of differentially methylated regions, ultimately advancing biological understanding and translational potential of epigenetics.

The thesis of Wenbin Guo is approved.

Qing Zhou

Ying Nian Wu

Jingyi Li, Committee Chair

University of California, Los Angeles

2024

To Matteo and Jessica

Table of Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

First and foremost, I sincerely thank my bioinformatics advisor, Dr. Matteo Pellegrini, and my statistics advisor, Dr. Jingyi Jessica Li, for their invaluable support and mentorship throughout this journey. Their guidance has been instrumental in shaping this thesis. I am especially grateful to Jessica for fostering my appreciation of statistics and its transformative role in science and understanding the world. Her encouragement, dedication, and commitment to excellence have been pivotal to my academic growth and will continue to inspire me in my future endeavors.

I extend my heartfelt gratitude to my committee members, Dr. Qing Zhou and Dr. Yingnian Wu, for their teaching and guidance. Dr. Zhou's classes deepened my understanding of statistical models, which laid the foundation for Chapter 3. Dr. Wu's courses on machine learning inspired Chapter 4 and its applications in bioinformatics. I also appreciate the great teachers in UCLA's Statistics and Biostatistics departments. Special acknowledgment goes to the second-order Taylor expansion and Jensen's inequality for simplifying complex theorems and making them approachable.

I want to thank the UCLA Statistics Department and Graduate Division for fostering a supportive academic environment, which enabled me to pursue additional statistical training alongside my PhD studies. I would also like to thank my statistics friends and collaborators for their insightful discussions and constructive feedback. Their expertise and support have been invaluable and made this journey more rewarding and enjoyable.

Last but not least, I am deeply grateful to my family and friends for their unwavering support and encouragement throughout every high and low of this journey, and to myself for embracing challenges, taking risks, and exploring new fields with curiosity and resilience.

Chapter 1

Introduction

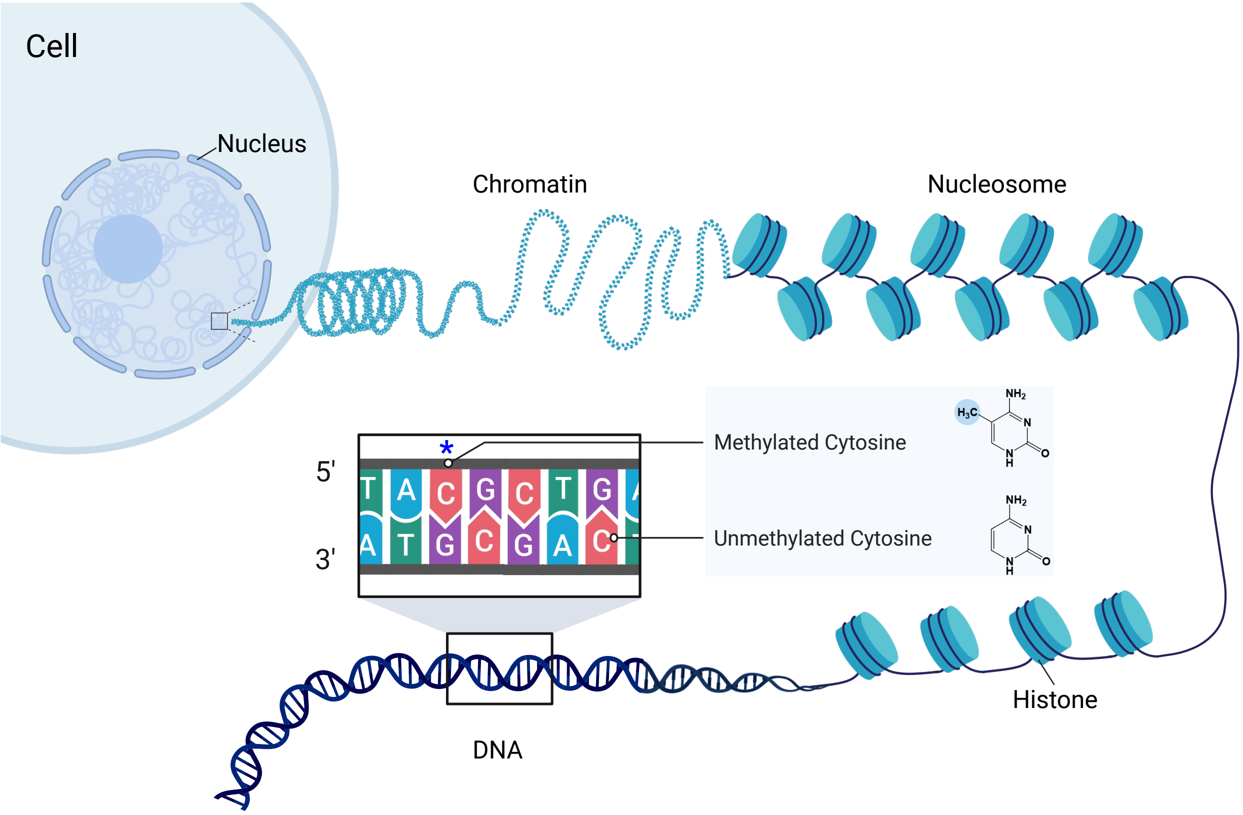
1.1 Background

1.1.1 DNA methylation

DNA methylation is a fundamental epigenetic modification involving the addition of a methyl group to the fifth carbon of the cytosine base, forming 5-methylcytosine (5-mC)(Figure 1.1). In mammals, this modification primarily occurs at CpG sites [1], where a cytosine is immediately followed by a guanine in the DNA sequence. The methylation states at these sites collectively define DNA methylation patterns, which are established and maintained by a family of enzymes known as DNA methyltransferases (DNMTs) [2]. While generally stable, these patterns can undergo dynamic changes, enabling DNA methylation to regulate gene expression and modulate diverse biological processes without altering the underlying DNA sequence [3].

t the molecular level, DNA methylation can suppress gene expression by blocking transcription factor binding and modifying chromatin structure, particularly when it occurs in gene promoter regions [4]. At the cellular level, methylation patterns play a critical role in cell differentiation and development, regulating gene activity in a controllable manner across different cell types and developmental stages [5]. At the systemic level, aberrant methylation patterns are implicated in various diseases, including cancer [6], metabolic syndromes [7], and neurological disorders [8]. Understanding the role of DNA methylation in these processes and diseases is crucial for advancing our knowledge of fundamental biological mechanisms, as well as improving disease diagnosis and therapeutic interventions.

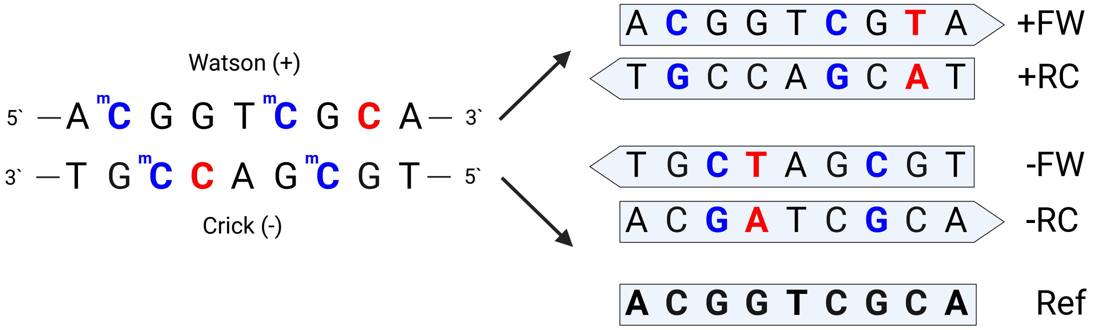
Figure 1.1: DNA methylation overview.



1.1.2 Bisulfite sequencing

Currently, the gold standard technology for DNA methylation profiling is bisulfite sequencing [9]. In this process, DNA fragments undergo bisulfite treatment, where unmethylated cytosines (C) are converted to uracil (U) and subsequently read as thymine (T) during sequencing. In contrast, methylated cytosines (mC) remain unchanged and are still read as cytosine (C) (Figure 1.2). The basic unit of bisulfite sequencing data is called a sequencing read, which is a readout of a short DNA fragment sequence, typically several hundred base pairs long. After aligning the observed reads to the reference genome, the methylation states of the cytosines can be determined by comparing the observed base on the read to the reference base on the genome, where a C to T base change[1] indicates no methylation, while C remaining as C indicates methylation. This approach allows for precisely measuring methylation states at single-base resolution across the genome.

Figure 1.2: Bisulfite sequencing for DNA methylation detection. The left panel shows a DNA fragment with methylated cytosine (mC, blue color) and unmethylated cytosine (C, red color). The right panel shows the readout of the DNA fragment in bisulfite sequencing, with the reference genome shown at the bottom for comparison and base changes highlighted in red.

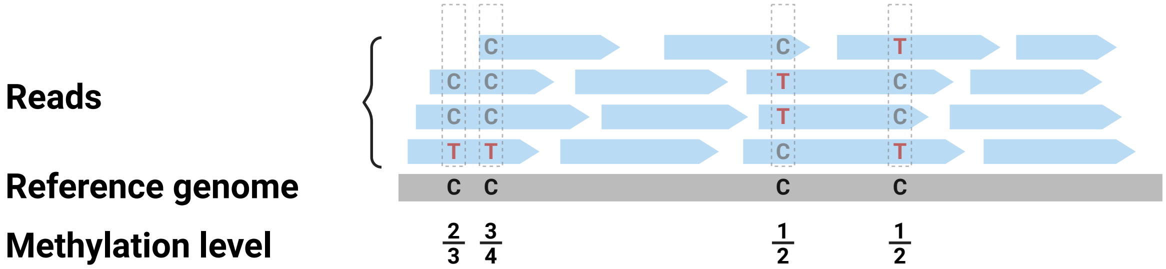


At the per-read level, each methylable site has two states: unmethylated (encoded as 0) or methylated (encoded as 1). At the population level, where multiple cells and copies of chromosomes are sequenced, each methylable site t has a methylation level m_t, which measures the methylation rate of site t in the cell population and takes a continuous value from 0 to 1. A value of 0 means no cell is methylated at site t, and a value of 1 means every cell is methylated at site t (see Figure 1.3). By calculating the ratio of reads supporting methylation (i.e., reads with no base change at site t) to the total number of reads covering the site, the methylation level m_t can be estimated by:

$$\hat{m_t} = \frac{\#\text{ methylated reads covering site } t}{\#\text{ total reads covering site } t} \in [0, 1] \tag{1.1}$$

[1] or G to A base change if the read maps to the reverse complementary strand. For illustration convenience, we only discuss the forward strand and C to T base change; the reverse complementary strand and G to A base change can be solved in a similar fashion.

Figure 1.3: Methylation level quantification. In conventional DNA methylation analysis, bisulfite sequencing reads (blue arrows) are aligned to the reference genome (grey segment). Vertical dashed boxes show the read count summary for each methylable site on the genome, with their methylation levels annotated at the bottom.

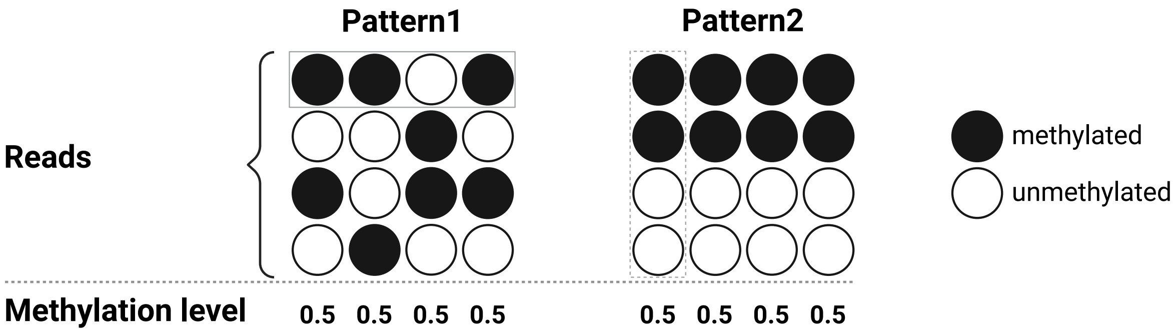


This data summary can be viewed as a form of data compression, where hundreds of millions of sequencing reads from a biological sample are represented as a long numerical vector named the methylation profile. Each element of this vector is a methylable site on the genome, with the corresponding value indicating its estimated methylation level. Since DNA methylation records environmental exposures and plays a crucial role in disease progression, current research on DNA methylation biomarkers primarily focuses on associating methylation profiles with various physiological traits and diseases. Such profiles are also widely used as features to predict individual traits and health outcomes, such as aging [10], disease status [7], and patient survival [11].

1.2 Motivation

While the data summary procedure is simple and efficient, it treats each site as an independent feature and considers only the first-order statistic (mean). In contrast, higher-order statistics, such as interactions between sites (site-site dependencies), are ignored. As illustrated in the toy example (Figure 1.4), Pattern 1 and Pattern 2 display distinct methylation patterns across sequencing reads; however, their summarized methylation levels appear identical. The site-site dependency ignored by the data summary procedure in Pattern 2 is also known as co-methylation, where methylable sites proximal in genome locations are more likely to share the same methylation state. Such dependency has been demonstrated in previous studies [12, 13] and validated in real WGBS data analysis (Figure 2.1).

Figure 1.4: Methylation pattern on sequencing reads. The horizontal solid box shows the methylation pattern on a sequencing read; the vertical dashed box shows the methylation level quantification for each site. This toy example highlights the information loss in conventional data summary procedures where distinct methylation patterns yield identical summarized methylation levels.



In the past few years, there has been an interest in reversing the data compression process for synthetic data generation, where the read-level methylation patterns are generated *in silico* using the site-level methylation profile. This interest leads to a handful of synthetic data generation tools named bisulfite sequencing simulators such as Sherman [14], BSSim [15], MethylFastQ [16], BSBolt [17] (Table 1.1). These simulators generate synthetic bisulfite sequencing reads with known ground truth, such as reads' origin and sites' true methylation levels, thus serving as an important approach for both new bioinformatics tool development and existing tools benchmark, such as bisulfite read aligners, SNP callers, etc. However, almost all simulators generate methylation patterns on a read using independent Bernoulli models (section 2.2) where each site is treated independently, neglecting their coordination. WGBSSuit [18] claimed the ability to simulate dependency in methylation levels and generate read summary counts for each site. However, it focuses on site-level summary statistics and does not account for generating the detailed methylation patterns or site-site dependencies within individual reads. To this end, accurately modeling site-site dependency and generating realistic methylation patterns at the read level remains an open challenge, which this thesis seeks to address.

Table 1.1: Summary of existing bisulfite sequencing simulator

Simulator	Year	Generative model	Output	Link
Sherman	2011	Independent Bernoulli	Reads	[14]
BSSim	2014	Independent Bernoulli	Reads	[15]
MethylFASTQ	2019	Independent Bernoulli	Reads	[16]
BSBolt	2020	Independent Bernoulli	Reads	[17]
WGBSSuit	2015	Binomial	Read counts	[18]
pWGBSimla	2020	Binomial	Read counts	[19]

1.3 Relevance/Impact

Modeling site-site dependencies has multiple benefits and significant implications. First, after modeling the site-site dependency, it can be seamlessly integrated into bisulfite sequencing simulators to generate realistic synthetic data, allowing bioinformatics tools to be evaluated under more representative and practical conditions. Additionally, it holds the potential to deepen our understanding of methylation mechanisms by investigating the factors driving the dependencies captured by the model. Furthermore, modeling site-site dependencies can support a variety of applications in DNA methylome analysis, including missing value imputation, differential methylated region detection, and cellular composition deconvolution. Advancements in these areas will enhance analytical precision, ultimately leading to more reliable and insightful interpretations of methylation data.

1. Missing value imputation

Assuming site t has a true methylation level m_t and reads are independently sampled from this level, the total methylation count follows a binomial distribution with probability m_t. The methylation level estimate, $\hat{m}_t$, defined as the ratio of methylated counts to total counts N_t, is an unbiased estimator of m_t with variance:

$$\mathbb{V}\text{ar}(\hat{m}_t) = \frac{m_t(1 - m_t)}{N_t}$$

For sites with low sequencing depth (N_t is small), the estimate suffers from high uncertainty due to sampling randomness. Conventional methylation analyses often exclude low-depth sites, treating them as missing. This approach reduces sampling noise at the cost of introducing missing data. Most missing value imputation methods, such as blocked k-nearest-neighbor (KNN), typically disregard site-site dependencies and assign equal weights to all sites. However, adjacent sites often provide more relevant information about the missing values due to local dependency, which suggests an opportunity for improvement. By incorporating spatial dependencies, weighted imputation methods can leverage information from neighboring sites more effectively, enhancing imputation accuracy.

2. Differential methylated region analysis

Differentially methylated region (DMR) analysis is a critical approach for studying DNA methylome variations across conditions, such as health versus disease. The goal is to identify genomic regions with significant methylation differences between target and background groups. Typically, the genome is divided into short windows, and statistical tests (e.g., t-test, Mann-Whitney U test, Kolmogorov-Smirnov test) are applied to compare methylation levels between groups. However, these methods often treat individual sites and regions as independent features, disregarding their dependencies. By modeling site-site dependencies, joint probabilities across sites can be derived, enabling more sophisticated modeling and hypothesis-testing strategies. Incorporating these dependencies has the potential to improve the power and accuracy of DMR analyses, facilitating the detection of biologically meaningful signals.

3. Cellular composition deconvolution

Due to the high cost and low coverage of current single-cell bisulfite sequencing technologies, most data are generated from bulk samples, where mixed cell populations are sequenced together. Consequently, DNA methylation measurements represent composite signals, obscuring specific cell-type information critical for disease diagnosis and mechanistic studies. Inferring cell-type abundance from the composite measurement is a central task in DNA methylome analysis. Most existing deconvolution methods rely on first-order summary statistics, such as site-level methylation levels. However, recent studies have demonstrated that methylation patterns across sequencing reads also contain valuable information about cell type and clonal identity [13, 20]. Incorporating site-site patterns into deconvolution methods can improve accuracy by leveraging additional cell-type-specific features, particularly under low sequencing depth where the first-order summary statistics are noisy. Furthermore, understanding these dependencies could shed light on epigenetic regulatory mechanisms and aid in identifying novel cell types or subtypes.

1.4 Structure of the thesis

This thesis aims to model site-site dependencies in DNA methylation patterns of sequencing reads using two complementary approaches: probabilistic modeling with heterogeneous Hidden Markov Models (HMM) and deep learning-based sequence modeling with bidirectional Long Short-Term Memory (LSTM) networks. The thesis is organized into the following chapters:

- chapter 1: **Introduction**

Chapter 1 provides an overview of DNA methylation and bisulfite sequencing technologies, summarizes current analytical approaches, and highlights the motivation for modeling site-site dependencies.

- chapter 2: **Problem formulation**

  Chapter 2 introduces the mathematical notations for the problem setting, describes the commonly used independent Bernoulli model that overlooks site-site dependencies, and explores site-site dependency in real WGBS data, setting the stage for the subsequent chapters.

- chapter 3: **Modeling site-site dependency using heterogeneous HMM**:

  Chapter 3 presents an explicit modeling approach for site-site dependencies using heterogeneous Hidden Markov Model (HMM). It provides a detailed overview of the model framework, its implementation, and its connections to alternative approaches. The chapter evaluates the performance of the heterogeneous HMM on both simulated and real data, comparing it to the classical homogeneous HMM.

- chapter 4: **Modeling site-site dependency using bidirectional LSTM**

  Chapter 4 investigates an implicit modeling approach using bidirectional Long Short-Term Memory (BiLSTM) networks. It details the neural network architecture and how it captures site-site dependencies without explicitly defining interaction rules. The chapter compares the performance of BiLSTM to the HMM approaches, emphasizing the strengths and flexibilities of neural networks in capturing complex, non-linear dependencies and preserving target marginal distributions.

- chapter 5: **Conclusion and Discussion**

  Chapter 5 summarizes the key findings from the two modeling approaches. It also outlines future work directions and provides an outlook for potential epigenetic research applications.

Chapter 2

Problem Formulation

2.1 Notations

First, Let the genome length be L, the total number of methylable sites on the genome be K, and the total number of sequencing reads be N. We define

$$\{\mathcal{S}_s\}_{s=1}^{K} : \text{ the collection of methylable sites on the genome}$$

$$\{\mathcal{R}_r\}_{r=1}^{N} : \text{ the collection of bisulfite sequencing reads}$$

Due to the short length of sequencing reads, each read $\mathcal{R}_r$ only spans a small subset of consecutive methylable sites on the genome. Let T_r represent the total number of methylable sites covered by $\mathcal{R}_r$; the methylable sites on the read can be represented as:

$$Y_r = (y_{[r,1]}, y_{[r,2]}, \ldots, y_{[r,T_r]}) : \text{observed methylation states on read } \mathcal{R}_r \text{ where } y_{[r,t]} \in \{0, 1\}$$

$$C_r = (c_{[r,1]}, c_{[r,2]}, \ldots, c_{[r,T_r]}) : \text{genomic coordinates of sites where } c_{[r,t]} \in \{1, \ldots, L\}$$

$$M_r = (m_{[r,1]}, m_{[r,1]}, \ldots, m_{[r,T_r]}) : \text{true methylation levels of sites where } m_{[r,t]} \in [0, 1]$$

$$D_r = (d_{[r,1]}, d_{[r,1]}, \ldots, d_{[r,T_r]}) : \text{genomic distance between the current site } t \text{ and last site } t-1$$

$$H_r = (h_{[r,1]}, h_{[r,1]}, \ldots, h_{[r,T_r]}) : \text{genomic context embedding of each methylable sites}$$

where $t = 1, \ldots, T_r$, indexing the t-th methylable site (also referred to as site t) on the read. For illustration convenience, we will drop subscript r when analyzing a single read.

2.2 Independent Bernoulli model

The problem we aim to address is to generate a realistic methylation pattern Y of a read based on available features. Pioneering work predominantly utilized the independent Bernoulli model where the sites are treated as independent from each other (Figure 2.3, Model 1), and the methylation state of site t, as denoted by y_t, is modeled as a random variable drawn from a site-specific Bernoulli distribution.

$$y_t \sim \text{Bern}(m_t) \tag{2.1}$$

or

$$P(y_t) = m_t^{y_t}(1 - m_t)^{1-y_t} \quad \text{for } y_t \in \{0, 1\} \tag{2.2}$$

For a read consisting of T consecutive methylable sites, the methylation pattern is represented as $Y = (y_1, y_2, \ldots, y_T) \in \{0, 1\}^T$. Under the site independence assumption, the probability of observing such a methylation pattern is the product of observing each site.

$$P(Y = (y_1, y_2, \ldots, y_T)) = \prod_{t=1}^{T} m_t^{y_t}(1 - m_t)^{1-y_t} \tag{2.3}$$

In practice, the true methylation level m_t is plugged in by the methylation level estimates $\hat{m}_t$. Such a generative model ignores the higher-order relationship among sites (site-site dependency). As a result, the generated methylation patterns are more randomized and exhibit reduced concordance between adjacent sites compared to real data, as indicated by the preliminary analysis (Figure 2.2); This reduced realism arises because neglecting site-site dependency effectively decouples adjacent sites, diminishing the spatial correlation inherent in the actual data, ultimately making the synthetic data less realistic.

2.3 Site-site dependency in WGBS data

To demonstrate site-site dependency in real bisulfite sequencing data, we utilized Whole Genome Bisulfite Sequencing (WGBS) data from the PGP-UK project [21]. This dataset provides high-depth DNA methylation measurement at single-base resolution across the genome, enabling the study of dependencies between adjacent CpG sites. By analyzing sequencing reads covering consecutive CpG sites, we investigated whether methylation states of nearby sites are correlated and quantified the extent of this dependency.

2.3.1 Data collection and processing

Data collection

The WGBS data used in this study originates from a human blood sample in the PGP-UK project and is publicly available through the European Nucleotide Archive (Project ID: PRJEB17529, Sample Accession ID: ERR2359938). The raw sequencing data consists of approximately 660 million paired-end reads, each 150 bp in length. We downloaded this data to the UCLA IDRE hoffman2 cluster, where subsequent data processing and analysis steps were performed.

Data processing

The raw sequencing data underwent quality control, including adapter trimming and removal of low-quality bases, using fastp (version 0.23.2) [22]. The processed reads were aligned to the high coverage GRCh38 reference genome, obtained from the 1000 Genomes Project [23], using BSBolt (version 1.5.0) [17]. Methylation levels for each CpG site were calculated following BSBolt's standard pipeline, which includes extracting methylation calls from aligned reads and summarizing methylation proportions at each CpG position. All software parameters were set to their default values unless otherwise specified.

Methylation pattern extraction

The methylation patterns of individual reads were determined by comparing the read sequences to the reference genome at the aligned positions. To ensure robust measurements, we focused on regions with at least 20 read counts spanning over 500 bp. Only reads mapped to chromosome 21 on the Watson strand and covering at least four CpG sites were included for further analysis. After filtering, approximately 16,000 reads were retained for modeling and analysis. The processed data and analysis pipeline are publicly available on GitHub: https://github.com/wbvguo/Site-site_dependency.

2.3.2 Dependency metrics

After extracting the methylation patterns of individual reads, site-site dependency between adjacent sites can be quantified by summarizing the methylation states from reads spanning both sites. Consider two adjacent sites $\mathcal{S}_t$ and $\mathcal{S}_{t+1}$ on the genome, the observed methylation states for each site on a given read are denoted as $y_t, y_{t+1} \in \{0, 1\}$. The possible methylation state combination $y_t y_{t+1} \in \{00, 01, 10, 11\}$ are then summarized and counted across all reads covering both sites, resulting in the contingency table shown in Table 2.1.

Table 2.1: Contingency table summarizing methylation state pairs for adjacent sites across reads.

		y_{t+1}		
		0	1	Sum
y_t	0	n_{00}	n_{01}	$n_{0\cdot}$
	1	n_{10}	n_{11}	$n_{1\cdot}$
	Sum	$n_{\cdot 0}$	$n_{\cdot 1}$	n

The joint probability of observing methylation state pairs can be estimated as

$$\mathbb{P}(y_t = i, y_{t+1} = j) = \frac{n_{ij}}{n} = \hat{p}_{ij} \quad \text{for } i, j \in \{0, 1\}$$

where $n = n_{00} + n_{01} + n_{10} + n_{11}$ is the total number of reads covering both sites. And the marginal probabilities of methylation state for each site are:

$$\mathbb{P}(y_t = i) = \frac{\sum_{j=0}^{1} n_{ij}}{n} \quad \text{and} \quad \mathbb{P}(y_{t+1} = j) = \frac{\sum_{i=0}^{1} n_{ij}}{n}$$

Using these probabilities, we define the following metrics to assess site-site dependency:

1. Probability of Same States

The probability of adjacent sites sharing the same state is given by:

$$\mathbb{P}(y_t = y_{t+1}) = \frac{n_{00} + n_{11}}{n} = \hat{p}_{00} + \hat{p}_{11} \tag{2.4}$$

This metric quantifies the state concordance of adjacent sites being either both unmethylated (00) or both methylated (11).

2. Entropy

Entropy measures the randomness of the methylation patterns across reads:

$$\mathbf{H} = -\sum_{y_t, y_{t+1} \in \{0,1\}} \mathbb{P}(y_t, y_{t+1}) \log \mathbb{P}(y_t, y_{t+1}). \tag{2.5}$$

Lower entropy indicates stronger site-site dependency, as fewer patterns dominate the distribution.

3. Mutual Information

Mutual information (MI) quantifies the dependency between two random variables. In our case, we used it to evaluate how much information about one site's state is gained by knowing the state of the other:

$$\mathbf{MI}(y_t, y_{t+1}) = \sum_{y_t, y_{t+1} \in \{0,1\}} \mathbb{P}(y_t, y_{t+1}) \log \left(\frac{\mathbb{P}(y_t, y_{t+1})}{\mathbb{P}(y_t)\mathbb{P}(y_{t+1})} \right). \tag{2.6}$$

Particularly, when the sites are independent, MI is zero.

4. State Correlation

State correlation provides a normalized measure of the linear relationship between adjacent sites' states. It is calculated as:

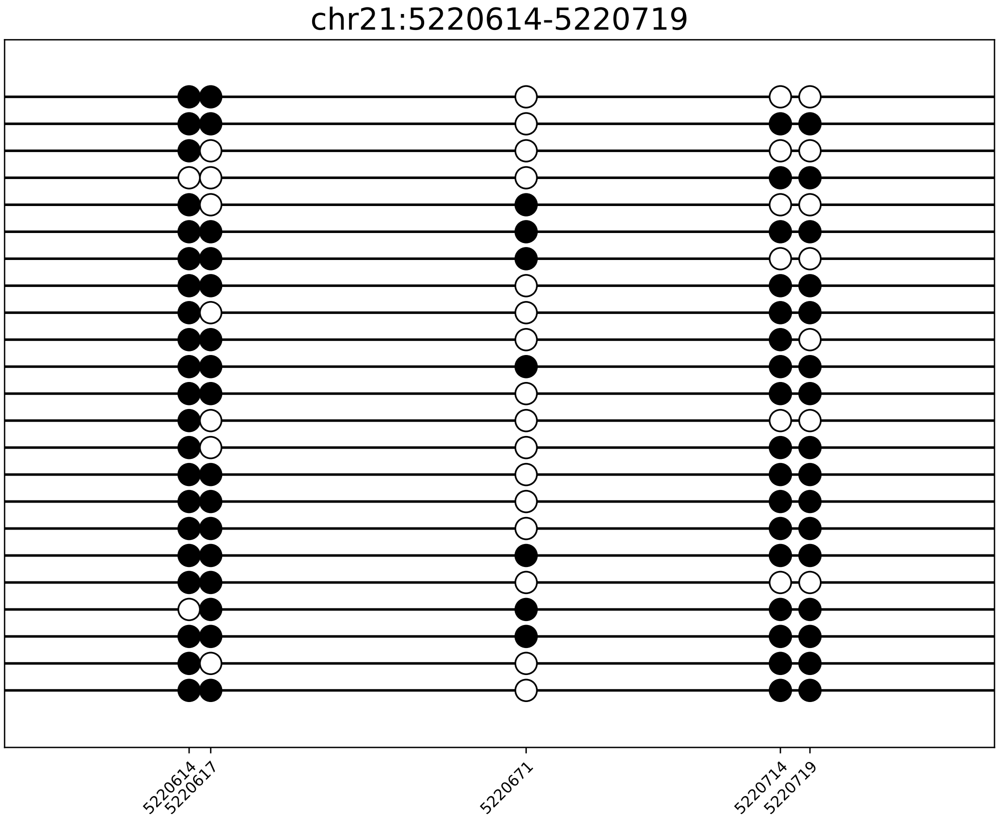
$$\mathrm{Corr}(y_t, y_{t+1}) = \frac{\mathbb{P}(y_t = 1, y_{t+1} = 1) - \mathbb{P}(y_t = 1) \cdot \mathbb{P}(y_{t+1} = 1)}{\sqrt{\mathbb{P}(y_t = 1)(1 - \mathbb{P}(y_t = 1)) \cdot \mathbb{P}(y_{t+1} = 1)(1 - \mathbb{P}(y_{t+1} = 1))}}. \tag{2.7}$$

These metrics enable the quantitative evaluation of site-site dependency across reads and provide a framework for comparing real and simulated data. Models that do not account for site-site dependency, such as the Bernoulli model, typically exhibit higher entropy, lower mutual information, and reduced state correlation compared to data with inherent site-site dependency, such as real WGBS data.

2.3.3 Dependency examination in WGBS data

As an example of dependency examination in WGBS data, Figure 2.1 illustrates how site-site dependency varies with genomic distance. By summarizing read counts of state pairs for adjacent sites and calculating dependency metrics, we observed that closer adjacent sites exhibit stronger dependency, evidenced by higher probabilities of sharing the same state (e.g., 0.70 vs. 0.39) and lower entropy (e.g., 1.30 vs. 1.73). In contrast, longer distances correspond to weaker dependencies, characterized by increased entropy and reduced probabilities of shared states. Mutual information and state correlation also decline with increasing distance, reflecting a loss of dependency, though their values may vary due to sparse state configurations. These findings suggest that spatial proximity influences the methylation patterns observed in sequencing reads from WGBS data.

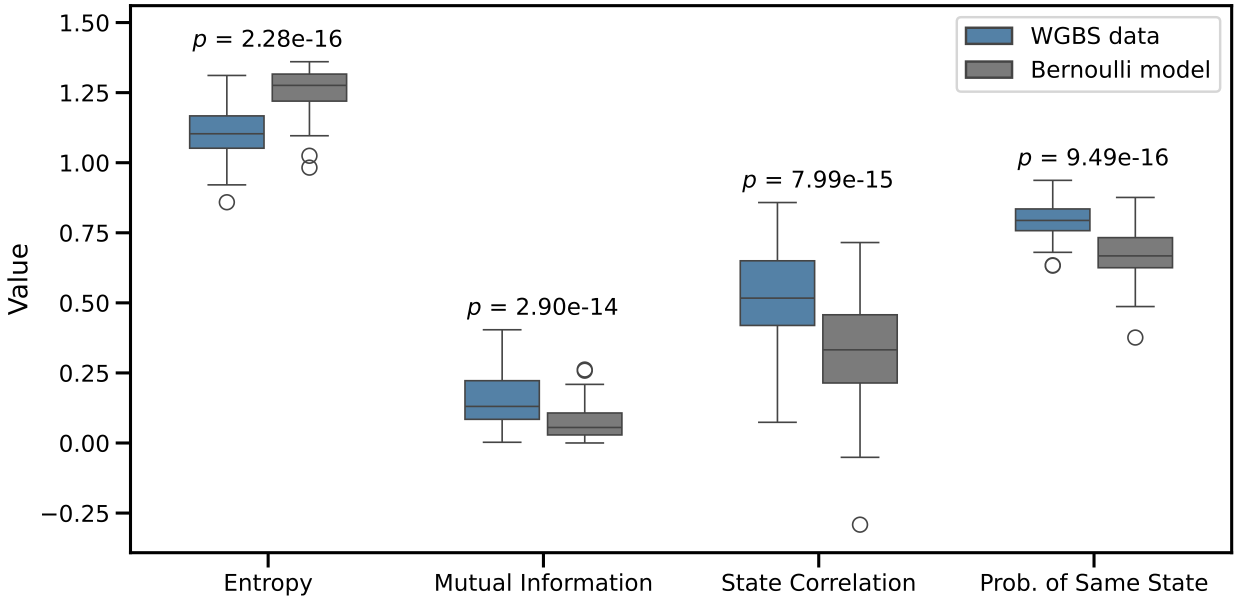
Figure 2.1: Distance-related site-site dependency in WGBS data.



# 00	1	6	4	6
# 01	1	1	12	0
# 10	6	10	2	1
# 11	15	6	5	16
Entropy	1.30	1.73	1.71	1.07
Prob. of Same State	0.70	0.52	0.39	0.96
Mutual Information	0.01	0.04	0.00	0.65
State Correlation	0.13	0.23	-0.04	0.90

The observed methylation pattern on a read, where proximal sites tend to share the same methylation state, could partially be explained by the similarity of marginal methylation propensities. This phenomenon arises when closely located genomic sites exhibit similar inherent methylation potentials, resulting in state similarity driven by individual site characteristics rather than genuine site-site dependency. To disentangle these factors, we conducted a simulation using the independent Bernoulli model that preserves the marginal methylation levels of each site but assumes no dependency between adjacent sites. Dependency metrics of adjacent sites, including entropy, mutual information, state correlation, and the probability of the same state, were calculated and compared between the simulated and real WGBS data. As shown in Figure 2.2, the Bernoulli model exhibits significantly higher entropy, lower mutual information, and reduced state correlation compared to the real data, indicating weaker dependency. This discrepancy provides compelling evidence that the observed methylation patterns in real data are not solely driven by local methylation propensities but reflect genuine site-site dependency, which the Bernoulli model fails to capture.

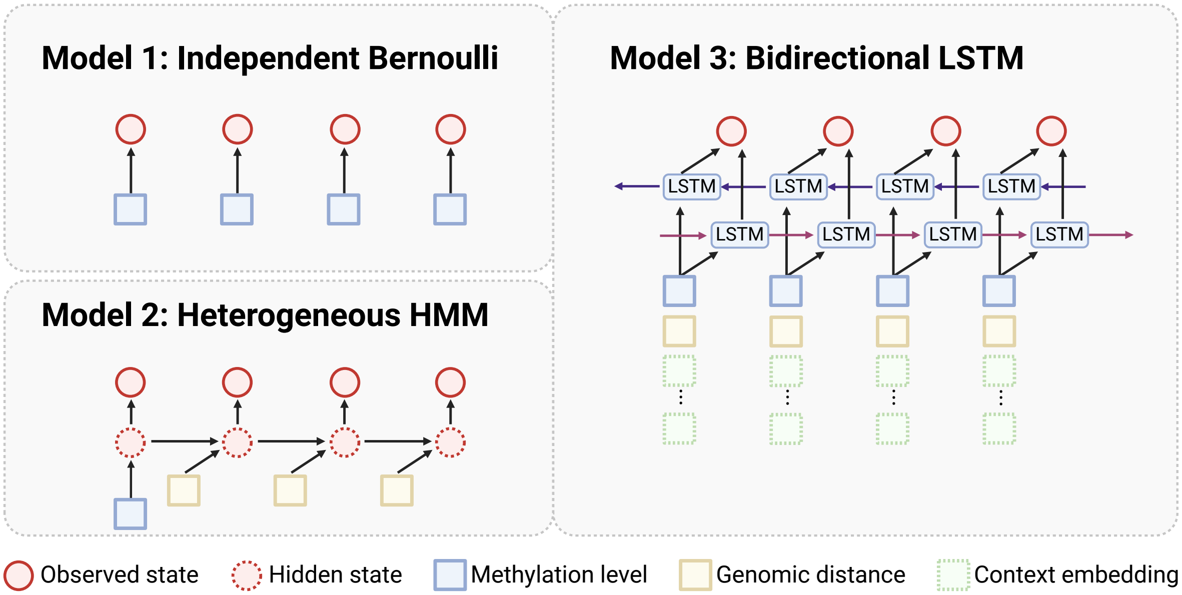
Figure 2.2: Comparison of dependency metrics between real WGBS data and the independent Bernoulli model predictions.



2.4 Site-site dependency modeling strategies

To address the limitations of site-independent models, we proposed two distinct approaches to capture site-site dependency, shown as model 2 and model 3 in Figure 2.3. These methods aim to generate more realistic methylation patterns by incorporating other genomic features, which are ignored in simplified models like the independent Bernoulli model. By exploring both statistical and machine learning paradigms, we provide complementary strategies for addressing the complexity of site-site relationships.

Figure 2.3: Three strategies for DNA methylation pattern modeling on sequencing reads.


Model 1: Independent Bernoulli
Model 2: Heterogeneous HMM
Model 3: Bidirectional LSTM
LSTM
Observed state
Hidden state
Methylation level
Genomic distance
Context embedding

chapter 3 introduces a statistical modeling approach, where we develop an explicit parametric model to directly quantify and account for site-site dependency. This method emphasizes interpretability, enabling a clear view of the underlying dependency structure. In chapter 4, we shift focus to a neural network-based approach, leveraging recurrent neural networks to implicitly learn site-site dependency patterns from data. Unlike the parametric model, this approach does not require predefined assumptions and instead relies on the network's capacity to uncover dependencies during training. Through simulations and real data analyses, we evaluate the effectiveness of these approaches, demonstrating their ability to generate biologically meaningful methylation patterns while addressing the shortcomings of simplified, site-independent models.

Chapter 3

Heterogeneous Hidden Markov Model

Markov models, particularly Hidden Markov Models (HMMs), have long been integral to statistical modeling in bioinformatics. Their applications span a broad spectrum of tasks, including multiple sequence alignment [24], gene discovery [25], epigenome segmentation [26], regulatory motif and binding site prediction [27], among others. Traditional (or classical) Markov models typically assume homogeneous state transition probabilities, meaning that the system's dynamics remain constant over time and that the transition probabilities between states are fixed. While this assumption simplifies the modeling process, it may fail to capture the time-varying nature of certain real-world processes. In terms of modeling site-to-site dependencies in genomic sequences, such an assumption can be overly restrictive because state dependencies between adjacent sites may vary due to external factors such as base context, genomic distance, etc.

In this chapter, we introduce the heterogeneous Hidden Markov Model (heterogeneous HMM), an extension of the classical HMM framework that allows transition probabilities to vary as a function of genomic distance between adjacent sites. This extension provides a more flexible approach to modeling site-to-site dependencies, effectively addressing the inherent heterogeneity found in epigenomic data [28]. By accommodating variable state transitions, the heterogeneous HMM can capture more complex dependency structures than those possible with classical HMMs (also referred to as homogeneous HMM in the following context).

We begin by introducing the structure of heterogeneous HMM and discussing its relationship to other statistical models. We then derive the parameter estimation procedure based on the maximum likelihood approach and solve it using the Expectation-Maximization

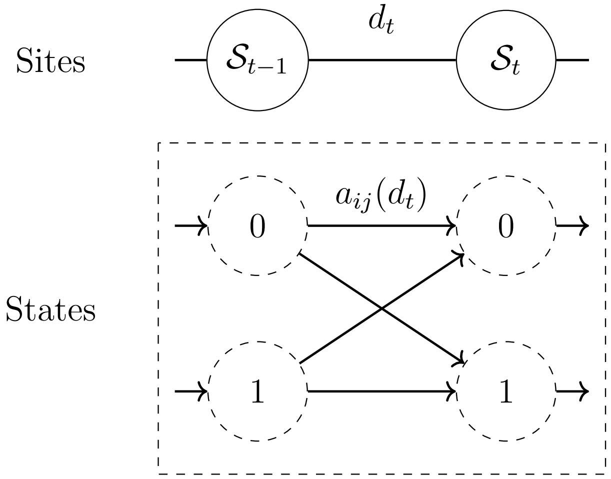
(EM) algorithm. Through simulation studies, we assess the performance of the heterogeneous HMM compared to the homogeneous HMM. Finally, we apply our model to real data, demonstrating its practical utility in capturing complex dependency structures in biological sequences.

3.1 Model setup

Similar to homogeneous HMM, the heterogeneous HMM consists of two major parts: the state transition model (subsection 3.1.1) and the emission model (subsection 3.1.2).

3.1.1 State transition model

Figure 3.1: Schematic of the state transition model.



Consider a pair of adjacent sites $\mathcal{S}_{t-1}$ and $\mathcal{S}_t$ with genomic distance d_t (Figure 3.1), the transition probability from last state $z_{t-1} = i$ to current state $z_t = j$ is defined as $a_{ij}(d_t)$. Due to that each site on a read can take two states: unmethylated (0) and methylated (1); the transition probability matrix $\mathbf{A}(d_t)$ is a 2×2 matrix and can be written as

$$\mathbf{A}(d_t) = \Big[a_{ij}(d_t)\Big]_{2\times 2} = \begin{pmatrix} a_{00}(d_t) & a_{01}(d_t) \\ a_{10}(d_t) & a_{11}(d_t) \end{pmatrix} \tag{3.1}$$

Here, the rows stand for the states of site $\mathcal{S}_{t-1}$, and columns are the states of site $\mathcal{S}_t$. Inspired by previous work in genome segmentation [29], we hypothesized the transition matrix can be partitioned into two parts: the distance-irrelevant part and the distance-dependent part. The former is a constant matrix, while the latter takes the form of a constant matrix times a decay factor controlled by the genomic distance d_t. The decomposition formula is given by:

$$\mathbf{A}(d_t) = \mathbf{A}_1 + \phi(d_t)\mathbf{A}_2 \tag{3.2}$$

where

$$\mathbf{A}_1 = \begin{array}{c} \\ 0 \\ 1 \end{array}\!\!\begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1-p_1 & p_1 \\ p_2 & 1-p_2 \end{pmatrix} \end{array} \quad \text{and} \quad \mathbf{A}_2 = \begin{array}{c} \\ 0 \\ 1 \end{array}\!\!\begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} p_1 & -p_1 \\ -p_2 & p2 \end{pmatrix} \end{array} \tag{3.3}$$

with rows and columns representing the states of site $\mathcal{S}_{t-1}$ and site $\mathcal{S}_t$, respectively, and the decay factor $\phi(d_t) \in [0, 1]$ is a monotonic decreasing function of d_t. For simplicity, we model this decay factor using a generalized sigmoid function[1], where $\phi(d_t) = \text{sigmoid}(w_0 + w_1 d_t)$, with unknown parameters $w_0 \in \mathbb{R}$ and $w_1 \in \mathbb{R}^-$. Thus,

$$\begin{cases} a_{00}(d_t) = 1 - p_1 + \phi(d_t)p_1 = 1 - p_1 \cdot \text{sigmoid}(-w_0 - w_1 d_t) \\ a_{01}(d_t) = p_1 - \phi(d_t)p_1 = p_1 \cdot \text{sigmoid}(-w_0 - w_1 d_t) \\ a_{10}(d_t) = p_2 - \phi(d_t)p_2 = p_2 \cdot \text{sigmoid}(-w_0 - w_1 d_t) \\ a_{11}(d_t) = 1 - p_2 + \phi(d_t)p_2 = 1 - p_2 \cdot \text{sigmoid}(-w_0 - w_1 d_t) \end{cases} \tag{3.4}$$

[1] We also tested the truncated exponential function $\phi(d_t) = \min(e^{w_0 + w_1 d_t}, 1)$, which yielded less accurate parameter estimates in simulation. Thus, we chose the generalized sigmoid function throughout this work.

From Equation 3.2, it's easy to notice that when

$$\begin{aligned} \phi(d_t) \to 1, \quad \mathbf{A}(d_t) \to \mathbf{I}_{2\times 2} \\ \phi(d_t) \to 0, \quad \mathbf{A}(d_t) \to \mathbf{A}_1 \end{aligned} \tag{3.5}$$

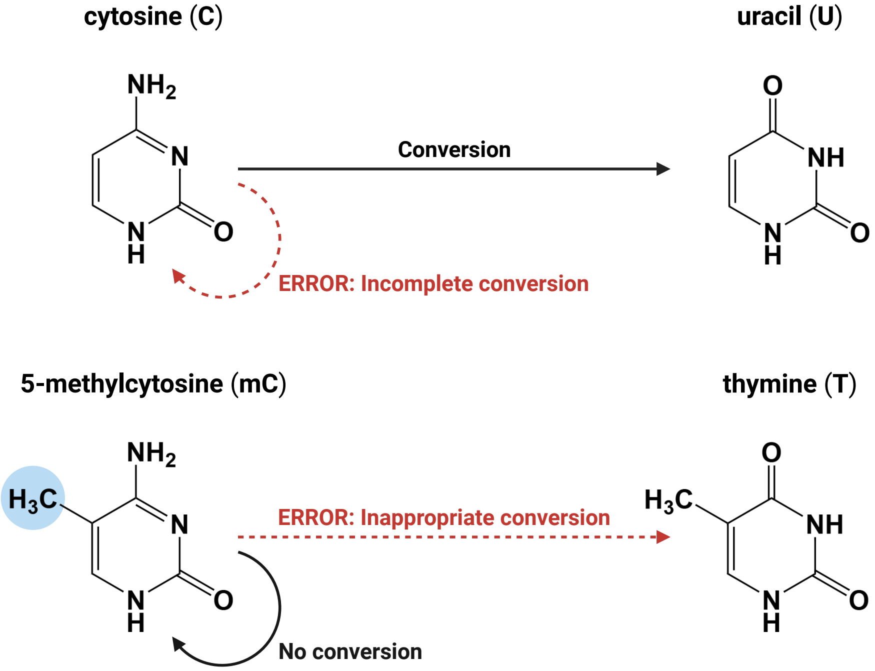
Specifically, when $\phi(d_t) \to 1$, the transition matrix will approximate to a 2×2 identity matrix which makes the methylation state to be consecutive, aligning with previous observations of co-methylation in closely-located adjacent sites [12]. On the other hand, when $d_t > 0$ is large, $\phi(d_t) \to 0$, the contribution of $\mathbf{A}_2$ to the transition matrix can be ignored, making the adjacent states to be less concordant. This numerical property makes biological sense where the longer the distance between two sites, the less dependency two sites will have as the increased distance changes the biophysical and thermodynamic landscape of the surrounding microenvironment, therefore decoupling the sites.

3.1.2 Emission model

To determine the methylation state at each site on a DNA fragment, the fragment undergoes bisulfite conversion followed by sequencing, allowing the methylation state to be read from the sequencing output (namely the sequencing reads). Bisulfite conversion typically converts unmethylated cytosines (C) into uracil (U), which are subsequently replaced and read as thymine (T) during sequencing, while methylated cytosines remain unchanged. However, due to the stochastic nature of the molecular processes involved, each process may introduce errors. Specifically, two types of bisulfite conversion errors can occur [30] (Figure 3.2).

- Incomplete conversion: unmethylated C fails to convert to U.
- Inappropriate conversion: methylated C is erroneously converted to T.

Figure 3.2: Two types of bisulfite conversion errors.



Beyond bisulfite conversion errors, additional errors can arise during the sequencing process. These errors occur when the sequencing system incorrectly identifies a base, commonly referred to as sequencing errors [31]. With advancements in sequencing technologies, the error rates have become remarkably low and can generally be assumed as uniform across different nucleotide types. Given the characteristics of DNA methylation and bisulfite sequencing, we consider three typical sequencing error scenarios for methylable sites, assuming a uniform sequencing error rate of ϵ

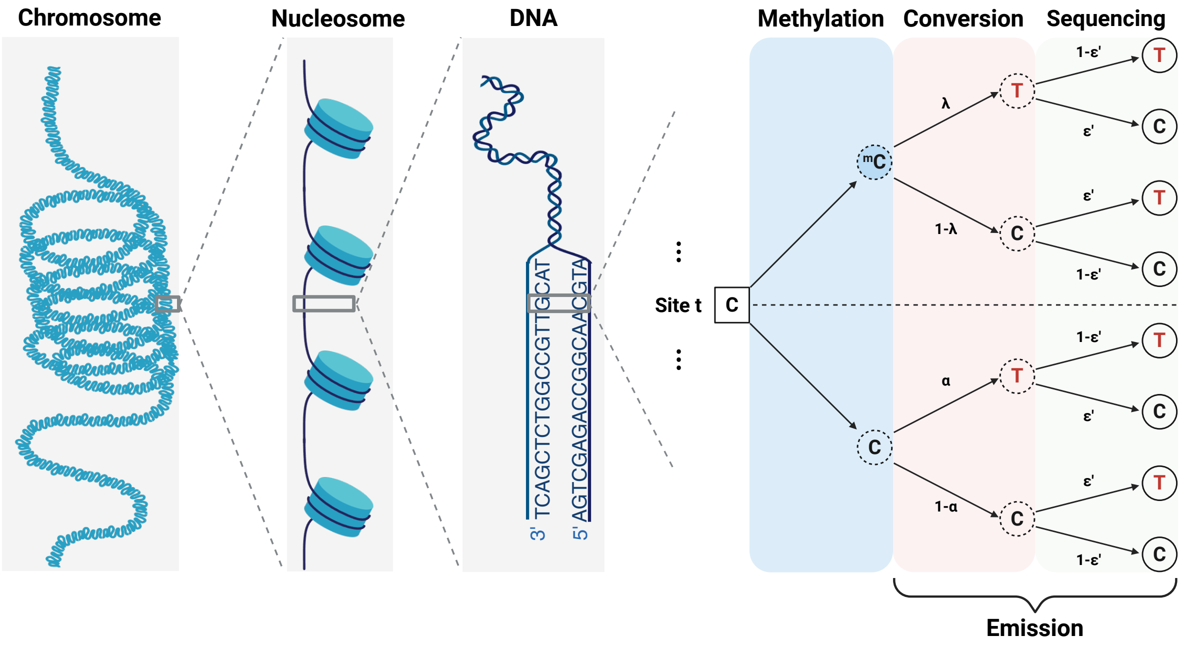
- C is observed as C with probability $1 - \epsilon$.
- C is observed as T with probability $\frac{1}{3}\epsilon$.
- C is observed as other bases (A or G) with probability $\frac{2}{3}\epsilon$.

The second point arises because when a sequencing error occurs at base C, there are three possible incorrect observations (A, G, or T), each equally likely. In data collection and processing, observations corresponding to non-T bases (A or G) can be easily recognized as sequencing errors and typically filtered out. Consequently, the third scenario is excluded from the analysis data. After normalization to ensure the probabilities sum to 1, the effective error rate for observing C as T, denoted as ϵ' is

$$\epsilon' = \frac{\frac{1}{3}\epsilon}{\frac{1}{3}\epsilon + 1 - \epsilon} = \frac{\epsilon}{3 - 2\epsilon} \tag{3.6}$$

Collectively, let α represent the bisulfite conversion success rate, where unmethylated cytosine is converted to T (also referred to as the conversion rate in some literature), λ the inappropriate conversion rate, and ϵ' the effective sequencing error rate. These three parameters can be treated as constants for a given experiment with consistent physicochemical conditions. Since the chemical processes (conversion and sequencing) are independent from the biological process (methylation), each step can be modeled as an independent sampling from a Bernoulli distribution. Given the stepwise nature of these processes, the entire emission process can be described using a hierarchical model, as shown in Figure 3.3.

Figure 3.3: Schematic of the emission model.



Let y_t denote the observed methylation state at site t, and let z_t represent the corresponding true but unobserved methylation state. The emission probability $b_j(k)$ is defined as the probability of observing $y_t = k$ given that the true hidden state is $z_t = j$. Given the binary nature of the hidden states—unmethylated (0) and methylated (1)—and the observed symbols—C-to-T change (0) and C remains unchanged (1)—the emission probability matrix $\mathbf{B} = \Big[b_j(k)\Big]$ is a 2×2 matrix, which can be written as:

$$\mathbf{B} = \begin{array}{ccc} & 0 & 1 \\ 0 & b_0(0) & b_0(1) \\ 1 & b_1(0) & b_1(1) \end{array} \quad \text{where} \begin{cases} b_0(0) = \alpha(1-\epsilon') + (1-\alpha)\epsilon' = 1 - p_3 \\ b_0(1) = \alpha\epsilon' + (1-\alpha)(1-\epsilon') = p_3 \\ b_1(0) = \lambda(1-\epsilon') + (1-\lambda)\epsilon' = p_4 \\ b_1(1) = \lambda\epsilon' + (1-\lambda)(1-\epsilon') = 1 - p_4 \end{cases} \tag{3.7}$$

Here the rows represent the hidden states, and the columns correspond to the observed symbols. It is evident that when the parameters $\alpha, \lambda,$ and ϵ are held constant, the emission probability matrix becomes fixed and can be reparameterized using two parameters p_3, p_4.

3.1.3 Connections to other models

Compared to homogeneous HMM, the heterogeneous HMM introduces a modified definition of transition probabilities by allowing them to be modulated based on the genomic distance between adjacent sites. This added flexibility enables the model to capture more complex dependencies inherent in genomic data. Notably, the proposed heterogeneous HMM encompasses several well-known models as special cases, highlighting its versatility and connections to existing methodologies. Recognizing these special cases is crucial for both theoretical understanding and practical implementation.

Reduction to the homogeneous HMM

Firstly, from the transition perspective, when the decay function $\phi(d_t)$ is a constant and takes value c (i.e., does not vary with genomic distance), the heterogeneous HMM degenerates to the homogeneous HMM. In this scenario, the transition probability matrix simplifies to:

$$\mathbf{A}(d_t) = \mathbf{A}_1 + \phi(d_t)\mathbf{A}_2 = \mathbf{A}_1 + c\mathbf{A}_2 = \text{constant}$$

As a result, the transition probabilities become independent from genomic distance, leading to time-invariant system dynamics. Therefore, the homogeneous HMM can be viewed as a specific instance within our more general heterogeneous HMM framework.

Approximation to a heterogeneous Markov Chain

Secondly, from the emission perspective, current experimental protocols exhibit very high bisulfite conversion rates ($\alpha > 0.995$) [32] and low inappropriate conversion rates $\lambda < 0.01$ [30]. Additionally, advanced sequencing technologies have very low sequencing error rates ($\epsilon < 0.003$) [31]. As a result, the emission probability matrix $\mathbf{B}$ is approximately an identity matrix $\mathbf{I}_{2\times2}$, implying that the observed data closely reflect the true underlying states.

In other words, the emission probabilities are nearly deterministic, allowing the hidden states to be treated as directly observed. This simplification effectively reduces the heterogeneous HMM to a heterogeneous Markov chain, where the primary focus shifts to modeling the variable transition probabilities between states rather than inferring unobserved states. Consequently, the problem simplifies to estimate the parameters of a heterogeneous Markov chain that maximize the likelihood of the observed state sequence.

3.2 Parameter estimation

Let us consider a collection of sequencing reads $\mathcal{R}_r$ for $r = 1, \ldots, N$, each covering T_r methylatable sites along the genome. For each read $\mathcal{R}_r$, we denote the t-th methylable site as site t and define:

- Observed methylation pattern Y_r: a vector $(y_{[r,1]}, y_{[r,2]}, \ldots, y_{[r,T_r]})$, where $y_{[r,t]}$ represents the observed methylation state at site t.

- True methylation pattern Z_r: a vector $(z_{[r,1]}, z_{[r,2]}, \ldots, z_{[r,T_r]})$, where $z_{[r,t]}$ denotes the unobserved true methylation state at site t.

- Adjacent genomic distance D_r: a vector $(d_{[r,1]}, d_{[r,2]}, \ldots, d_{[r,T_r]})$, where $d_{[r,t]}$ is the genomic distance between site t and the preceding site $t-1$ on read R_r. We set the $d_{[r,1]} = 0$ due to the lack of a preceding site for the initial site on the read.

The initial state probability $\pi_r = \mathbb{P}\left(z_{[r,1]}\right)$ for each read $\mathcal{R}_r$ is assumed to be known and given by the methylation level of the first methylable site on the read, serving as the starting point for the Markov process. By setting $d_{[r,1]} = 0$, we allow the first site's state to be governed by π_r, while subsequent states are determined by the heterogeneous Markov Chain, which depends on the preceding states and the genomic distances.

3.2.1 Maximum likelihood estimate (MLE) framework

Our objective is to estimate the unknown parameters $\boldsymbol{\theta} = (p_1, p_2, \omega_0, \omega_1, p_3, p_4)$ given the observed reads data $\{Y_r\}_{r=1}^N$ and covariate data $\{D_r\}_{r=1}^N$, while accounting for the unobserved true methylation patterns $\{Z_r\}_{r=1}^N$. To achieve this, we employ the maximum likelihood estimation approach. Since each read corresponds to a DNA fragment from a biological specimen, the reads can be treated as independent and identically distributed(i.i.d.) samples. Therefore, the likelihood of the entire dataset can be expressed as:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{P}(\{Y_r\}_{r=1}^N \mid \{D_r\}_{r=1}^N, \boldsymbol{\theta}) = \prod_{r=1}^{N} \mathbb{P}(Y_r \mid D_r, \boldsymbol{\theta}) = \prod_{r=1}^{N} \mathcal{L}_r(\boldsymbol{\theta}) \tag{3.8}$$

where $\mathcal{L}_r(\boldsymbol{\theta}) = \mathbb{P}(Y_r \mid D_r, \boldsymbol{\theta})$ represents the likelihood of the observed sequence Y_r from a single read $\mathcal{R}_r$. Since the likelihood of the full dataset factorizes into the product of individual likelihoods for each read (yielding an additive log-likelihood), we can streamline the parameter estimation process by first analyzing a single read sequence, then extend the procedure to the full dataset by aggregating the likelihoods across all reads. This approach simplifies parameter estimation without sacrificing generality.

For illustrative purposes, we omit the subscript r when analyzing a single read sequence. Specifically, we can write the observation data likelihood of a single read as

$$\mathbb{P}(Y \mid D, \boldsymbol{\theta}) = \sum_{Z} \mathbb{P}(Y, Z \mid D, \boldsymbol{\theta}) \tag{3.9}$$

with the sum taken over all possible methylation patterns Z for the given read. This marginalization step involves summing over an exponentially large number of possible hidden state configurations, which can be computationally prohibitive. An alternative approach is to assume that the hidden states are known and maximize the joint probability of Y and Z given parameters $\boldsymbol{\theta}$ and genomic distances D (also known as the complete data likelihood). According to the Markov assumption, this complete data likelihood can be factorized as follows:

$$\begin{aligned}\mathbb{P}(Y, Z \mid D, \boldsymbol{\theta}) =& \mathbb{P}(y_1, \ldots, y_T, z_1, \ldots, z_T \mid d_1, \ldots, d_T, \boldsymbol{\theta}) \\ =& \mathbb{P}(z_1)\, \mathbb{P}(y_1 \mid z_1)\, \mathbb{P}(z_2 \mid z_1, d_2)\, \mathbb{P}(y_2 \mid z_2) \cdots \mathbb{P}(z_T \mid z_{T-1}, d_T)\, \mathbb{P}(y_T \mid z_T) \\ =& \mathbb{P}(z_1)\, \mathbb{P}(y_1 \mid z_1) \prod_{t=2}^{T} \mathbb{P}(z_t \mid z_{t-1}, d_t)\, \mathbb{P}(y_t \mid z_t) \end{aligned} \tag{3.10}$$

Here, $\mathbb{P}(z_1)$ is the initial distribution of z_1, which is assumed to be known when the starting site is given. The probability functions $\mathbb{P}(z_t \mid z_{t-1}, d_t)$ and $\mathbb{P}(y_t \mid z_t)$ represent the transition probability from state z_{t-1} to z_t given genomic distance d_t, and the emission probability of observing y_t given hidden state z_t, respectively. Let us denote the previous hidden state as $z_{t-1} = i$, the current hidden state as $z_t = j$, and the current observation as $y_t = k$. Using the indicator variable $z_{tj} = I(z_t = j)$, we can express the transition and emission probability functions as:

$$\mathbb{P}(z_t \mid z_{t-1}, d_t) = \{a_{ij}(d_t)\}^{z_{(t-1)i} z_{tj}} \tag{3.11}$$

$$\mathbb{P}(y_t \mid z_t) = \{b_j(k)\}^{z_{tj}} \tag{3.12}$$

Thus, the complete data likelihood is

$$\mathbb{P}(Y, Z \mid D, \boldsymbol{\theta}) \propto \prod_{i=0}^{1}\prod_{j=0}^{1}\prod_{t=2}^{T}\{a_{ij}(d_t)\}^{z_{(t-1)i}z_{tj}} \times \prod_{j=0}^{1}\prod_{k=0}^{1}\prod_{t:y_t=k}\{b_j(k)\}^{z_{tj}} \tag{3.13}$$

Then

$$\begin{aligned}\log \mathbb{P}(Y, Z \mid D, \boldsymbol{\theta}) &= \sum_{i,j}\sum_{t=2}^{T} z_{(t-1)i}z_{tj}\log a_{ij}(d_t) + \sum_{j,k}\underbrace{\sum_{t:y_t=k} z_{tj}}_{D_{jk}}\log b_j(k) \\ &= \sum_{i,j}\sum_{t=2}^{T} z_{(t-1)i}z_{tj}\log a_{ij}(d_t) + \sum_{j}\left[\sum_{k=0}^{1} D_{jk}\log b_j(k)\right] \end{aligned} \tag{3.14}$$

It can be observed that the log of the complete data likelihood naturally decomposes into two distinct components: a distance-dependent transition component, governed by the parameters $p_1, p_2, \omega_0, \omega_1$, and a distance-independent emission component, controlled by p_3, p_4. Notably, when Y, Z is given, the sufficient statistics for the emission model parameters $b_j(k)$ can be directly computed as:

$$D_{jk} = \sum_{t:y_t=k} z_{tj} \tag{3.15}$$

which represents the number of times symbol k is emitted from state j. These sufficient statistics can be used to derive the Maximum Likelihood Estimator (MLE) for the emission model parameters. However, deriving an analytical MLE for the transition model parameters directly from the complete data log-likelihood proves to be complex. Instead, we can seek a numerical optimization approach to obtain a solution, as detailed in the subsequent section.

3.2.2 Expectation-Maximization (EM) algorithm

Equation 3.14 provides a likelihood framework for parameter estimation. Specifically, when Z is known, we seek to find $\hat{\boldsymbol{\theta}}$ such that it maximizes the complete data log-likelihood:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log \mathbb{P}(Y, Z \mid D, \boldsymbol{\theta}) \tag{3.16}$$

Because the true methylation states Z are unobserved, we employ the Expectation-Maximization (EM) algorithm to iteratively estimate $\boldsymbol{\theta}$ by maximizing the expected complete-data log-likelihood.

E-Step: Calculate the expectation of complete-data log-likelihood

In the E-step, we compute the expectation of the complete data log-likelihood with respect to the posterior distribution of the hidden states Z, condition on the observed data Y, the distance D, and the current parameter estimates $\boldsymbol{\theta}^{(m)}$. The expected log complete data likelihood is

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)}) =& \mathbb{E}\left[\log \mathbb{P}(Y, Z \mid D, \boldsymbol{\theta}) \mid Y, D, \boldsymbol{\theta}^{(m)}\right] \\ =& \sum_{i,j} \sum_{t=2}^{T} \mathbb{E}\left[z_{(t-1)i} z_{tj} \log a_{ij}(d_t) \mid Y, D, \boldsymbol{\theta}^{(m)}\right] + \sum_{j,k} \sum_{t:y_t=k} \mathbb{E}\left[z_{tj} \log b_j(k) \mid Y, D, \boldsymbol{\theta}^{(m)}\right] \\ =& \sum_{i,j} \sum_{t=2}^{T} \mathbb{E}\left[z_{(t-1)i} z_{tj} \mid Y, D, \boldsymbol{\theta}^{(m)}\right] \log a_{ij}(d_t) + \sum_{j,k} \sum_{t:y_t=k} \mathbb{E}\left(z_{tj} \mid Y, D, \boldsymbol{\theta}^{(m)}\right) \log b_j(k) \end{aligned} \tag{3.17}$$

Thus, given model parameters $\boldsymbol{\theta} = \boldsymbol{\theta}^{(m)}$ and genomic distance covariates D, for each t and all i, j, we need to calculate:

$$\mathbb{P}\left(z_{t-1} = i, z_t = j \mid Y\right) \tag{3.18}$$

$$\mathbb{P}\left(z_t = j \mid Y\right) \tag{3.19}$$

Based on the conditional independence assumption of the Markov chain

$$\begin{aligned}
\mathbb{P}\left(z_t = j \mid Y\right) &\propto \mathbb{P}\left(Y, z_t = j\right) \\
&= \underbrace{\mathbb{P}\left(y_{1:t}, z_t = j\right)}_{\alpha_t(j)} \cdot \underbrace{\mathbb{P}\left(y_{(t+1):T} \mid z_t = j\right)}_{\beta_t(j)} \\
&= \alpha_t(j) \cdot \beta_t(j) && (3.20) \\
\Rightarrow \mathbb{P}\left(z_t = j \mid Y\right) &= \frac{\alpha_t(j) \cdot \beta_t(j)}{\sum_{j=0}^{1} \alpha_t(j) \cdot \beta_t(j)} := \gamma_t(j), \quad j = 0, 1 && (3.21)
\end{aligned}$$

On the other hand

$$\begin{aligned}
\mathbb{P}\left(z_{t-1} = i, z_t = j \mid Y\right) &\propto \mathbb{P}\left(Y, z_{t-1} = i, z_t = j\right) \\
&= \underbrace{\mathbb{P}\left(y_{1:(t-1)}, z_{t-1} = i\right)}_{\alpha_{t-1}(i)} \cdot \underbrace{\mathbb{P}\left(z_t = j \mid z_{t-1} = i\right)}_{a_{ij}(d_t)} \\
&\quad \times \underbrace{\mathbb{P}\left(y_t \mid z_t = j\right)}_{b_j(y_t)} \cdot \underbrace{\mathbb{P}\left(y_{(t+1):T} \mid z_t = j\right)}_{\beta_t(j)} \\
&= a_{ij}(d_t) \cdot b_j\left(y_t\right) \cdot \alpha_{t-1}(i) \cdot \beta_t(j) && (3.22)
\end{aligned}$$

By normalization, for all i, j,

$$\mathbb{P}\left(z_{t-1} = i, z_t = j \mid Y\right) = \frac{a_{ij}(d_t) \cdot b_j\left(y_t\right) \cdot \alpha_{t-1}(i) \cdot \beta_t(j)}{\sum_{i=0}^{1} \sum_{j=0}^{1} a_{ij}(d_t) \cdot b_j\left(y_t\right) \cdot \alpha_{t-1}(i) \cdot \beta_t(j)} := \xi_t(i, j) \qquad (3.23)$$

Recall that we denote $\alpha_{t-1}(i) = \mathbb{P}\left(y_{1:(t-1)}, z_{t-1} = i\right)$ and $\beta_t(j) = \mathbb{P}\left(y_{(t+1):T} \mid z_t = j\right)$, thus

$$\begin{aligned}
\alpha_t(j) &= \mathbb{P}\left(y_{1:t}, z_t = j\right) \\
&= \sum_{i=0}^{1} \mathbb{P}\left(y_{1:t}, z_t = j, z_{t-1} = i\right) \\
&= \sum_{i=0}^{1} \mathbb{P}\left(y_{1:(t-1)}, y_t, z_t = j, z_{t-1} = i\right) \\
&= \sum_{i=0}^{1} \mathbb{P}\left(y_{1:(t-1)}, z_{t-1} = i\right) \cdot \mathbb{P}\left(y_t \mid z_t = j\right) \cdot \mathbb{P}\left(z_t = j \mid z_{t-1} = i\right) \\
&= \sum_{i=0}^{1} \alpha_{t-1}(i) \cdot b_j(y_t) \cdot a_{ij}(d_t) \qquad (3.24)
\end{aligned}$$

$$\begin{aligned}
\beta_{t-1}(i) &= \mathbb{P}\left(y_{t:T} \mid z_{t-1} = i\right) \\
&= \sum_{j=0}^{1} \mathbb{P}\left(y_{t:T}, z_t = j \mid z_{t-1} = i\right) \\
&= \sum_{j=0}^{1} \mathbb{P}\left(y_{(t+1):T}, y_t, z_t = j \mid z_{t-1} = i\right) \\
&= \sum_{j=0}^{1} \mathbb{P}\left(y_{(t+1):T} \mid z_t = j\right) \cdot \mathbb{P}\left(y_t \mid z_t = j\right) \cdot \mathbb{P}\left(z_t = j \mid z_{t-1} = i\right) \\
&= \sum_{j=0}^{1} \beta_t(j) \cdot b_j(y_t) \cdot a_{ij}(d_t) \qquad (3.25)
\end{aligned}$$

The connection between $\alpha_{t-1}(i)$ and $\alpha_t(j)$, and $\beta_{t-1}(i)$ and $\beta_t(j)$ suggest we can calculate the values efficiently via recursive computing. Specifically, we can use the following forward algorithm (Algorithm 1) and backward algorithm (Algorithm 2) to compute both $\alpha_t(j)$ and $\beta_t(i)$ given $\theta^{(m)}$ and D; Once the $\alpha_t(j), \beta_t(i)$ is calculated, the conditional probability $\mathbb{P}\left(z_{t-1} = i, z_t = j \mid Y\right)$ and $\mathbb{P}\left(z_t = j \mid Y\right)$ can be obtained, so are their expectations.

Algorithm 1 Forward algorithm to calculate $\alpha_t(j)$ for all j and t

```
Initialize $\alpha_1(i) = \pi_i b_i^{(m)}(y_1)$ for $i = 0, 1$.
for $t = 2$ to $T$ do
    $\alpha_t(j) = b_j^{(m)}(y_t) \cdot \sum_{i=0}^{1} \alpha_{t-1}(i) \cdot a_{ij}^{(m)}(d_t), \quad j = 0, 1$
end for
```

Algorithm 2 Backward algorithm to calculate $\beta_t(i)$ for all i and t

```
Initialize $\beta_T(j) = 1$ for $j = 0, 1$.
for $t = T - 1$ to $1$ do
    $\beta_t(i) = \sum_{j=0}^{1} \beta_{t+1}(j) \cdot b_j^{(m)}(y_{t+1}) \cdot a_{ij}^{(m)}(d_t), \quad j = 0, 1$
end for
```

M-Step: find $\boldsymbol{\theta}^{(m+1)}$ by maximizing $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)})$

After calculating the expectations in the E-step, we can write function $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)})$ as

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)}) = \sum_{i,j} \sum_{t=2}^{T} \underbrace{\mathbb{E}\left[z_{(t-1)i} z_{tj} \mid Y, D, \theta^{(m)}\right]}_{c_{ij}^{(m)}(t)} \log a_{ij}(d_t) + \sum_{j,k} \underbrace{\sum_{t:y_t=k} \mathbb{E}\left(z_{tj} \mid Y, D, \boldsymbol{\theta}^{(m)}\right)}_{D_{jk}^{(m)}} \log b_j(k)$$

$$= \underbrace{\sum_{i,j} \sum_{t=2}^{T} c_{ij}^{(m)}(t) \log a_{ij}(d_t)}_{Q_1} + \underbrace{\sum_{j,k} D_{jk}^{(m)} \log b_j(k)}_{Q_2} \tag{3.26}$$

where $c_{ij}^{(m)}(t), D_{jk}^{(m)}$ are constants, and $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)})$ can be decomposed into 2 parts with Q_1 corresponds to the transition model and Q_2 corresponds to the emission model. Since these 2 models are independent in parameter space (Q_1 is controlled by p_1, p_2, w_0, w_1, Q_2 is controlled by $p_3, p4$), to maximize $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)})$, we can maximize Q_1 and Q_2 separately

As

$$
\begin{aligned}
Q_2 =& \sum_{j,k} D_{jk}^{(m)} \log b_j(k) \\
=& D_{00}^{(m)} \log b_0(0) + D_{01}^{(m)} \log b_0(1) + D_{10}^{(m)} \log b_1(0) + D_{11}^{(m)} \log b_1(1) \\
=& D_{00}^{(m)} \log(1-p_3) + D_{01}^{(m)} \log p_3 + D_{10}^{(m)} \log p_4 + D_{11}^{(m)} \log(1-p_4) \qquad (3.27)
\end{aligned}
$$

To maximize Q_2, we can take the partial derivative of Q_2 with respect to p_3 and p_4 and set it to 0; the resulting equation yields the Maximum likelihood estimator (MLE) for p_3, p_4 as

$$
p_3^{(m+1)} = \frac{D_{01}^{(m)}}{D_{00}^{(m)} + D_{01}^{(m)}} \qquad p_4^{(m+1)} = \frac{D_{10}^{(m)}}{D_{10}^{(m)} + D_{11}^{(m)}} \qquad (3.28)
$$

more generally,

$$
b_j(k)^{(m+1)} = \frac{D_{jk}^{(m)}}{D_{j\bullet}^{(m)}} \qquad (3.29)
$$

On the other hand, an analytical form of the estimates for p_1, p_2, w_0, w_1 are hard to derive.

$$
\begin{aligned}
Q_1 =& \sum_{t=2}^{T} c_{00}^{(m)}(t) \log a_{00}(d_t) + \sum_{t=2}^{T} c_{01}^{(m)}(t) \log a_{01}(d_t) + \sum_{t=2}^{T} c_{10}^{(m)}(t) \log a_{10}(d_t) + \sum_{t=2}^{T} c_{11}^{(m)}(t) \log a_{11}(d_t) \\
=& \sum_{t=2}^{T} c_{00}^{(m)}(t) \log\left(1 - p_1 \text{sigmoid}(-w_0 - w_1 d_t)\right) + \sum_{t=2}^{T} c_{01}^{(m)}(t) \log\left(p_1 \text{sigmoid}(-w_0 - w_1 d_t)\right) \\
&+ \sum_{t=2}^{T} c_{10}^{(m)}(t) \log\left(p_2 \text{sigmoid}(-w_0 - w_1 d_t)\right) + \sum_{t=2}^{T} c_{11}^{(m)}(t) \log\left(1 - p_2 \text{sigmoid}(-w_0 - w_1 d_t)\right) \\
=& f(p_1, p_2, w_0, w_1) \qquad (3.30)
\end{aligned}
$$

Actually, the sufficient statistics for p_1, p_2, w_0, and w_1 are difficult to obtain, and maximizing Q_1 is a more challenging problem as the transition probability is varying given genomic distance d_t. Therefore, we used a numerical optimization approach to obtain a solution

$$p_1^{(m+1)}, p_2^{(m+1)}, w_0^{(m+1)}, w_1^{(m+1)} = \arg \max_{\substack{p_1, p_2 \in [0,1] \\ w_0 \in \mathbb{R}, w_1 \in \mathbb{R}^-}} f(p_1, p_2, w_0, w_1) \tag{3.31}$$

In practice, We used the ‘L-BFGS-B‘ algorithm, a quasi-Newton method suitable for problems with bound constraints, to optimize the transition probabilities and decay parameters during the M-step. To speed up the convergence step of the M-step, we use the current estimate $(p_1^{(m)}, p_2^{(m)}, w_0^{(m)}, w_1^{(m)})$ of as the initial starting point for the optimization algorithm, which yields faster convergence speed during testing.

3.2.3 Algorithm summary

By combining the E-step and M-step, we can summarize the EM-algorithm as the following:

Algorithm 3 EM algorithm for heterogeneous HMM parameter estimation

```
Initialize parameters $\boldsymbol{\theta}^{(0)} = (p_1^{(0)}, p_2^{(0)}, w_0^{(0)}, w_1^{(0)}, p_3^{(0)}, p_4^{(0)})$ and construct $\mathbf{A}_1^{(0)}, \mathbf{A}_2^{(0)}, \mathbf{B}^{(0)}$
for $m = 1$ to max_iter do
    E-step:
    for each observation sequence $R_r$ with data $Y_r, D_r, \pi_r$ do
        Compute decay factors $\phi(d_t)^{(m)}$ for $t = 2, \ldots, T$ using the current $\boldsymbol{\theta}^{(m-1)}$.
        Compute transition probabilities $\mathbf{A}^{(m)}(d_t)$
        Compute forward probabilities $\alpha_t(i)$.
        Compute backward probabilities $\beta_t(i)$.
        Compute $\gamma_t(i)$ and $\xi_t(i, j)$.
    end for
    M-step:
    Update emission probabilities $\mathbf{B}^{(k)}$.
    Optimize $\mathbf{A}_1^{(k)}$, $\mathbf{A}_2^{(k)}$, and $\boldsymbol{w}^{(k)}$ by maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m-1)})$.
    Check for convergence:
    if converged then
        break
    end if
end for
Return $\boldsymbol{\theta}^{(m)}$
```

3.2.4 Implementation considerations

Convergence criteria

We implemented two types of convergence criteria based on the changes in (1) log-likelihood, or (2) parameter estimates between iterations. Specifically, the algorithm is considered to have converged if:

$$|\mathcal{L}(\boldsymbol{\theta}^{(m)}) - \mathcal{L}(\boldsymbol{\theta}^{(m-1)})| < \delta \quad \text{or} \quad \|\boldsymbol{\theta}^{(m)} - \boldsymbol{\theta}^{(m-1)}\|_\infty < \delta \tag{3.32}$$

where δ are small positive constants (default value: 10^{-5}). Through simulations, we found that the two criteria yield similar results. For consistency, the parameter convergence criterion was used throughout this thesis. Due to the non-convex nature of the optimization problem, the algorithm may converge to a local minimum. To mitigate this, we implemented a random start strategy, running the algorithm multiple times with different initializations and selecting the result with the highest likelihood. By default, we set n_starts= 20 for model training.

Numerical operations

We utilize numerical computing libraries such as NumPy and SciPy for optimized mathematical operations and numerical optimization routines. Besides, since the decay function involves the exponential functions, to prevent numerical underflow or overflow during computations, we apply clipping and add small constants where necessary. For example, we clip the input to the exponential function:

$$\lambda_t = \max(-\kappa, \min(w_1 d_t + w_0, \kappa)) \tag{3.33}$$

where κ is a large positive constant (e.g., 709 for double-precision floating-point numbers).

Code availability

A Python implementation of the heterogeneous HMM, heterogeneous Markov Chain, and homogeneous HMM can be accessed at the Github repository: https://github.com/wbvguo/Site-site_dependency.git

3.3 Model prediction and evaluation

3.3.1 Prediction

With predefined or estimated model parameters, the model can be used to generate new sequence predictions. Given an initial state distribution π and a sequence of distances D, the hidden and observed states can be generated using the following algorithm:

Algorithm 4 Sequence prediction using heterogeneous HMM

```
Initialize $z_1$ based on $\boldsymbol{\pi}$
for $t = 2$ to $T$ do
    Compute $\mathbf{A}_t = \mathbf{A}_1 + \phi(d_t)\mathbf{A}_2$
    Sample $z_t$ from $P(z_t \mid z_{t-1}, \mathbf{A}_t)$.
    Sample $y_t$ from $P(y_t \mid z_t, \mathbf{B})$.
end for
Return predicted sequence $Y = (y_1, ..., y_T)$ and hidden states $Z = (z_1, ..., z_T)$
```

3.3.2 Evaluation

We consider the following metrics to assess the model's performance in capturing the site-site dependency:

Root Mean Square Error

For synthetic data generated using a heterogeneous HMM with predefined parameters, we compute the Root Mean Square Error (RMSE) to quantify the deviation between the estimated parameters and the ground truth. This metric provides the parameter estimation accuracy, reflecting how closely the model captures the underlying generative process. The RMSE is given by:

$$\text{RMSE} = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta_i\right)^2} \tag{3.34}$$

where $\hat{\theta}_i$ represents the estimated parameter, θ_i is the corresponding true parameter, and n is the total number of parameters being evaluated.

Distance-related site-site dependency

After training the model, we also analyze the relationship between inter-site distance and various dependency measures, calculated based on the model's predicted sequences. Specifically, we compute the metrics defined in subsection 2.3.2 and evaluate how they vary with the distance between adjacent sites. This approach enables a detailed assessment of how well the model preserves the dependency structure presented in the synthetic or real-world WGBS data.

3.4 Simulation study

3.4.1 Synthetic data generation

To evaluate the performance of the heterogeneous HMM in parameter estimation, we conducted simulations based on Algorithm 4. With the given true model parameters $\boldsymbol{\theta}$, we generated synthetic datasets, each consisting of 100 observations (unless stated otherwise). Each observation consists of observed states Y, initial state probabilities π, and distances between adjacent sites D. The data generation details are outlined below:

1. **Initial State Probabilities**: For each observation sequence, the initial state probabilities π were sampled from a Dirichlet distribution $\pi \sim \text{Dirichlet}(\alpha = [0.5, 0.5])$, ensuring a diverse range of initial state distributions across sequences.

2. **Distance sequence**: The length of each sequence was randomly determined to range between 5 and 10 sites, and the distances between adjacent sites were sampled from a Uniform distribution $d_t \sim \text{Uniform}(1, 200)$, reflecting variability similar to the real-world data.

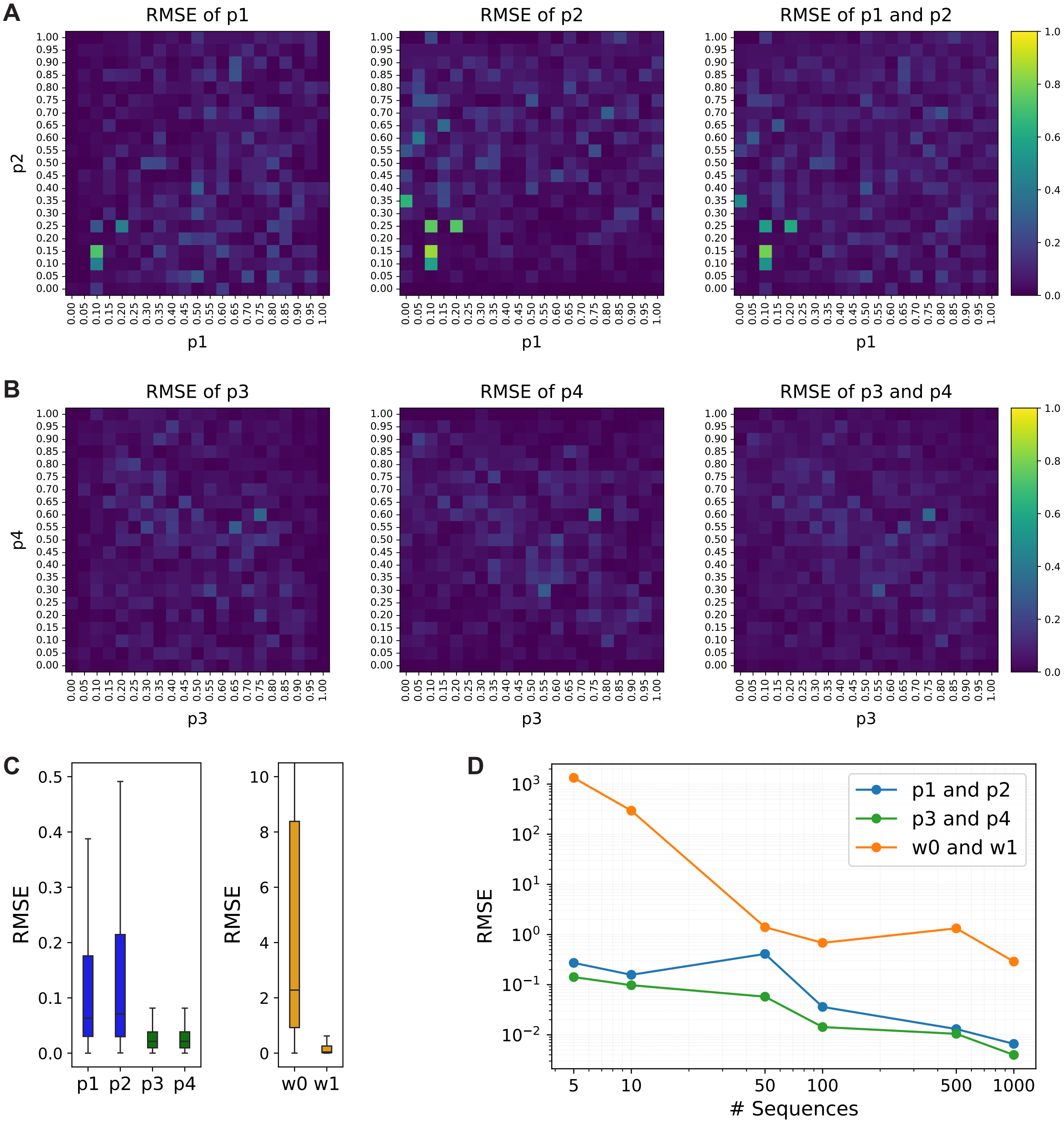
After generating the synthetic datasets, the heterogeneous HMM was fitted to each dataset to estimate the parameters. The model's performance in terms of fitting and parameter estimation accuracy was evaluated across various simulation settings.

3.4.2 Model performance evaluation

1. Parameter estimation accuracy

The parameter estimation performance of the heterogeneous HMM was evaluated by comparing the true and estimated model parameters under different simulation scenarios, as shown in Figure 3.4. In these simulations, we systematically varied three pairs of parameters: transition probabilities (p_1, p_2), emission probabilities (p_3, p_4), and decay parameters (w_0, w_1). For each scenario, one pair of parameters was varied while the remaining parameters were fixed for simulation to isolate their effects on estimation accuracy. We set $p_1 = p_2 = 0.5$, $p_3 = p_4 = 0.1$ and $w_0 = 5, w_1 = -0.05$ when they are not varied. After synthetic data generation and model fitting, we computed the RMSE for each parameter and parameter pair.

Figure 3.4: Parameter estimation accuracy of heterogeneous HMM in simulation. (A) RMSE for p_1, p_2 when varying the transition probabilities; (B) RMSE for p3 and p4 when varying the emission probabilities; (C) RMSE for the model parameters when the decay parameters; (D) RMSE as a function of the sample size.



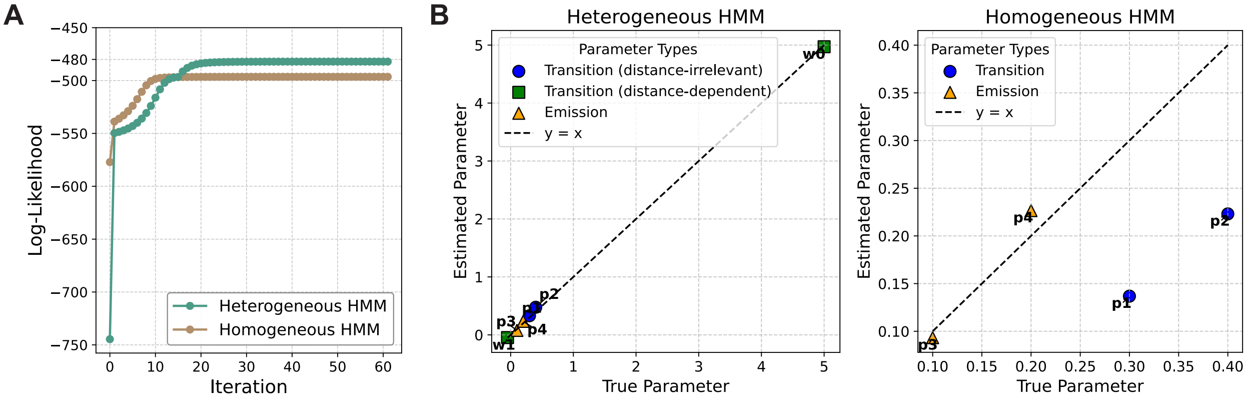
Panel A illustrates the RMSE for transition probabilities when these parameters are varied. While the estimates of p_1 and p_2 are generally accurate, high RMSE is occasionally observed, likely due to convergence issues during optimization. Panel B shows the RMSE for emission probabilities, which denotes p_3 and p_4 have smaller and less variable RMSE compared to that of p_1 and p_2, indicating they are easier to estimate. Moreover, the estimation accuracy of p_3 and p_4 slightly decreases when these probabilities greater than 0.25, meaning the emission model is less deterministic and functions more like a noise source in those cases. The overall trend reflects the model's robustness in estimating transition and emission probabilities.

Panel C provides additional insights into the challenges of estimating decay parameters. In this simulation, we randomly sampled the decay parameter w_0 from a Gaussian distribution $\mathcal{N}(5, 1)$ and w_1 from a half-normal distribution ($w_1 = -|w|$ where $w \sim N(0, 0.1)$). We repeated the sampling 500 times and generated a synthetic dataset for each pair of them. The parameter estimation results showed that w_0 and w_1 are harder to estimate, with significantly higher RMSE and greater variability than transition and emission probabilities, suggesting a potential lack of fitting with the current sample size. Panel D highlights that increasing the number of sequences reduces RMSE across all parameter categories, though the decay parameters improve more slowly. Overall, these findings suggest that the heterogeneous HMM is generally effective at accurately estimating model parameters.

2. Heterogeneous and homogeneous HMM comparison

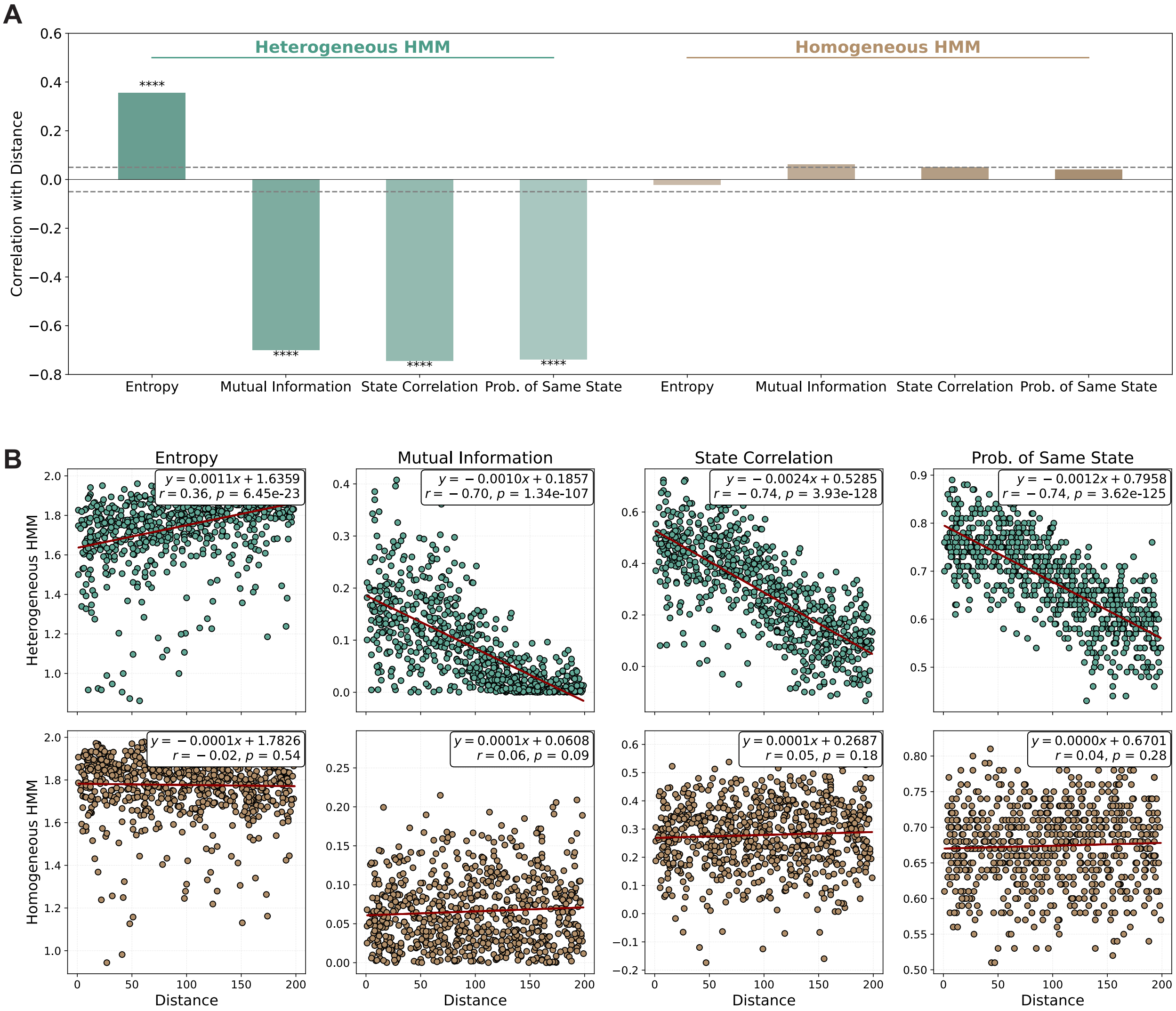
To emphasize the advantages of the heterogeneous HMM in capturing distance-related dependencies, we conducted a simulation, and compared its model fitting and parameter estimation performance to that of the homogeneous HMM. As shown in Figure 3.5, the heterogeneous HMM achieves a higher log-likelihood than the homogeneous HMM (Panel A), demonstrating its ability to capture the variability introduced by distance-dependent state transitions. In contrast, the homogeneous HMM, constrained by its assumption of constant transition probabilities, fails to account for this variability and converges to lower log-likelihood values. Panel B further illustrates that the heterogeneous HMM accurately estimates distance-dependent transition and emission probabilities, closely matching the true parameter values. While the homogeneous HMM estimates emission probabilities reasonably well, it systematically underestimates transition probabilities due to its inability to model distance dependency. These results underscore the heterogeneous HMM's superiority as a more flexible model for applications with varying transition probabilities.

Figure 3.5: Comparison of model fitting and parameter estimation between homogeneous and heterogeneous HMM in simulation.



Besides evaluating parameter estimation performance, we also assessed the ability of homogeneous and heterogeneous Hidden Markov Models (HMMs) to preserve distance-related dependencies in their predictions. For each adjacent site pair, the models generated 100 state predictions, and the correlation between distance and various site-site dependency measures—such as entropy, mutual information, state correlation, and the probability of sharing the same state—was analyzed. Figure 3.6 highlights the differences between the two models. Panel A demonstrates that the heterogeneous HMM consistently outperforms homogeneous HMM in preserving distance-related dependencies. Panel B further highlights this distinction through scatter plots. For the heterogeneous HMM, all measures show strong and consistent trends with distance, such as a significant negative correlation for mutual information and state correlation, indicating that the model correctly reflects decreasing similarity between adjacent sites as distance increases. Conversely, the homogeneous HMM exhibits weaker correlations, confirming that it fails to account for the effect of distance on state dependencies. These results reinforce the superiority of the heterogeneous HMM in modeling site-site dependencies, particularly in applications where distance plays a critical role in state transitions.

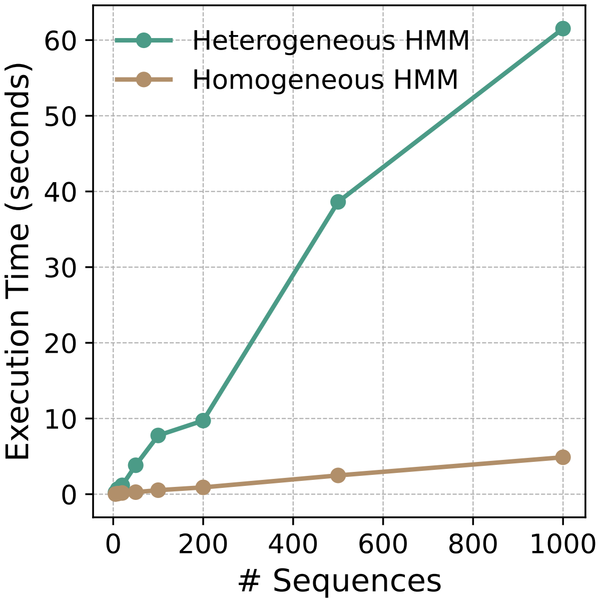
Figure 3.6: Comparison of distance-related dependency in homogeneous and heterogeneous HMM predictions.



3. Computational efficiency

To evaluate the computational speed of each method, we generated 1000 synthetic sequences and downsampled the sequence to construct observation datasets with sample sizes 5, 10, 50, 100, 200, 500, and 1000. For each dataset, we measured the computational time needed to fit the model on a PC with the following specifications (Intel 14700K, 64 GB RAM). The results show that the heterogeneous HMM requires more computation time than the homogeneous HMM (Figure 3.7). This difference is primarily due to the additional computations needed to calculate the distance-related transition probabilities and to perform the more complex M-step in the Expectation-Maximization (EM) algorithm for the heterogeneous model. In contrast, the homogeneous HMM achieves much faster execution times because its EM algorithm has a closed-form solution, which avoids iterative optimization during the M-step. Despite these differences in computational complexity, both methods exhibit a linear increase in execution time as the sample size grows, reflecting their scalability.

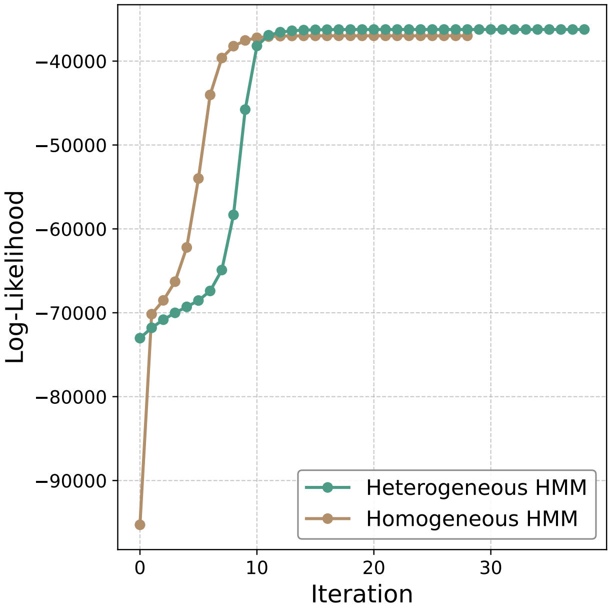
Figure 3.7: Comparison of computational efficiency for heterogeneous and homogeneous HMM.



3.5 Application to WGBS data

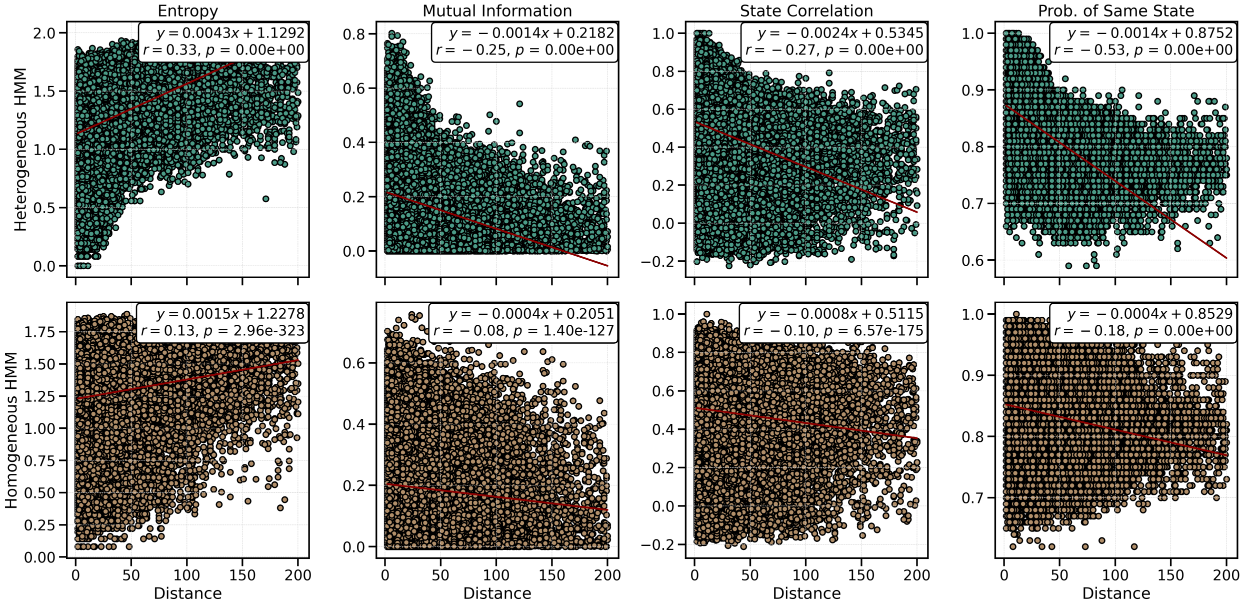
To assess the utility of heterogeneous HMM in modeling site-site dependencies in real WGBS data, we fitted both the heterogeneous and homogeneous HMM to WGBS data. Figure 3.8 illustrates the log-likelihood curves for both models, showing consistent increases as iterations progress. While the homogeneous HMM, which assumes uniform transition probabilities, converges more quickly, the heterogeneous HMM achieves slightly higher log-likelihood values, reflecting its flexibility in modeling distance-related transitions.

Figure 3.8: Comparison of model fitting for heterogeneous and homogeneous HMM in WGBS data.



We further assessed the heterogeneous HMM's ability to capture distance-related site-site dependencies by examining the relationships between the dependency metrics and distance, comparing it with the homogeneous HMM. As shown in Figure 3.9, the heterogeneous HMM demonstrates stronger associations between these metrics and distance. For example, entropy significantly increases with distance ($r = 0.33$), while mutual information, state correlation, and the probability of the same state decrease, reflecting the expected decay in dependency over larger distances. The homogeneous HMM, while capturing some dependency patterns, shows weaker correlations. These results highlight the superior capability of the heterogeneous HMM to model distance-related site-site dependencies, offering greater flexibility and accuracy compared to the homogeneous HMM in capturing the spatial variability inherent in WGBS data.

Figure 3.9: Comparison of dependency metrics for heterogeneous and homogeneous HMM data. Scatter plots showing the relationship between dependency metrics (Entropy, Mutual Information, State Correlation, and Probability of Same State) and genomic distance for the heterogeneous HMM (top) and homogeneous HMM (bottom). Red lines indicate fitted trends with correlation coefficients and p-values. The heterogeneous HMM captures stronger distance-related patterns, while the homogeneous HMM shows weaker associations.



Chapter 4

Bidirectional Long Short Term Memory

The parametric model presented in chapter 3 provides a statistical framework for capturing site-site dependencies in DNA methylation sequencing data. While simple and effective, it relies on assumptions that can be overly restrictive in certain contexts. Specifically, the model assumes that the true methylation pattern of a read is determined solely by initial state and transition probabilities, with dependencies varying only as a function of genomic distance. This assumption ignores other critical factors, such as the methylation potential of other sites on the read, sequence context, or motif pattern, which play key roles in shaping the methylation landscape [33]. Furthermore, the first-order Markov assumption—that each site depends only on the immediately preceding site's state—fails to account for long-range and bidirectional interactions, oversimplifying the spatial dependency structure of CpG sites. These limitations reduce the model's capability to capture the complex dependencies presented in real-world biological data.

To bypass these limitations, we explored an alternative approach based on bidirectional Long Short-Term Memory (BiLSTM) networks. Unlike the heterogeneous HMM, which explicitly encodes dependencies in the parameterized transition probabilities, BiLSTM is based on recurrent neural networks and implicitly learns sequential dependencies from the data. This capability allows BiLSTM networks to capture both short- and long-range interactions without relying on predefined, distance-based constraints, making them particularly well-suited for the complex dependency structures characteristic of genomic and epigenomic datasets. Additionally, by leveraging bidirectional processing, BiLSTM can incorporate contextual information from both preceding and following sequence features, further enhancing their ability to model intricate dependency patterns in genomic data.

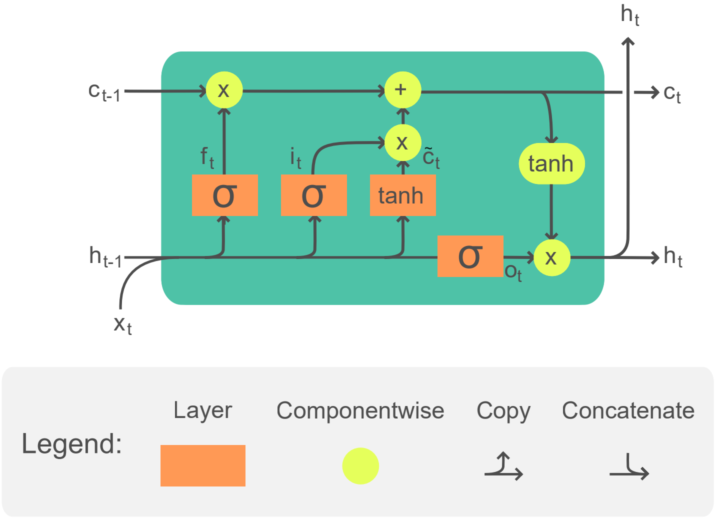
4.1 Method introduction

4.1.1 Recurrent neural network overview

Recurrent Neural Networks (RNNs) are a class of neural networks designed for modeling sequential data where each data point in the sequence is influenced by previous ones. They are widely used in tasks involving time-series prediction and sequence analysis due to their ability to capture temporal dependencies. However, standard RNNs struggle with long-range dependencies due to the vanishing and exploding gradient problem, which limits their ability to retain information over extended sequences.

To address these limitations, Hochreiter and Schmidhuber [34] introduced the Long Short-Term Memory (LSTM) network, which incorporates memory cells and gating mechanisms to effectively capture both short- and long-term dependencies. Each LSTM cell contains three gates: the **forget gate**, the **input gate**, and the **output gate** (Figure 4.1). These gates regulate the flow of information, controlling which information is retained, updated, or discarded, thereby allowing the network to maintain relevant information over long periods.

Figure 4.1: Long Short Term Memory (LSTM) architecture. This figure is adapted from Wikipedia with modifications.



Let's denote that the cell state (c_t) acts as the memory of the network, storing relevant information as it processes the sequence at time step t. The operations within an LSTM cell are mathematically defined as follows:

- **Input gate (i_t)**: The input gate determines which new information will be added to the cell state.

$$i_t = \sigma \left(W_i x_t + U_i h_{t-1} + b_i \right) \tag{4.1}$$

$$\tilde{c}_t = \tanh \left(W_c x_t + U_c h_{t-1} + b_c \right) \tag{4.2}$$

Cell Candidate ($\tilde{c}_t$): The candidate value for updating the cell state.

- **Forget gate (f_t)**: The forget gate decides what information should be retained or discarded from the cell state.

$$f_t = \sigma \left(W_f x_t + U_f h_{t-1} + b_f \right) \tag{4.3}$$

- **Cell state update (c_t):** The cell state is updated by combining the previous cell state and the candidate value, modulated by the forget and input gates.

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{4.4}$$

- **Output gate (o_t):** The output gate determines the output based on the updated cell state.

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \tag{4.5}$$

$$h_t = o_t \odot \tanh\left(c_t\right) \tag{4.6}$$

The hidden state (h_t) is passed to the next time step.

Where:

- x_t is the input at time t.
- h_{t-1} is the hidden state from the previous time step.
- c_{t-1} is the cell state from the previous time step.
- W_*, U_*, and b_* are weight matrices and biases to be learned.
- σ is the sigmoid activation function.
- tanh is the hyperbolic tangent activation function.
- $\odot$ denotes element-wise multiplication.

By incorporating these gating mechanisms and the cell state, LSTM networks effectively preserve long-term dependencies in sequential data, making them suitable for complex sequence modeling tasks such as language processing, speech recognition, and genomic sequence analysis.

4.1.2 Bidirectional Long Short-Term Memory network

While Long Short-Term Memory (LSTM) networks effectively model sequential dependencies, they process sequence information unidirectionally, typically from past to future. This limitation can hinder tasks requiring context from both directions. Bidirectional LSTM (BiLSTM) networks address this limitation by incorporating a second LSTM layer that processes the input sequence in reverse order. In a BiLSTM, the forward LSTM processes the input sequence $[x_1, x_2, \ldots, x_T]$, producing a sequence of hidden states $\{\overrightarrow{h_1}, \overrightarrow{h_2}, \ldots, \overrightarrow{h_T}\}$. Simultaneously, the backward LSTM processes the sequence in reverse order $[x_T, x_{T-1}, \ldots, x_1]$, producing $\{\overleftarrow{h_1}, \overleftarrow{h_2}, \ldots, \overleftarrow{h_T}\}$. The final hidden state at each time step is a concatenation of the forward and backward states:

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$$

This bidirectional structure allows BiLSTMs to capture dependencies from both preceding and succeeding contexts, enabling a more comprehensive understanding of sequential dependencies. The output of the BiLSTM at time t can be expressed as:

$$y_t = g(W_y \cdot h_t + b_y)$$

where y_t is the output at time t, W_y and b_y are the output weights and biases, and g is an activation function, such as softmax or sigmoid, depending on the task.

In genomic and epigenomic applications, dependencies can span diverse scales, influenced by *cis* and *trans* factors such as chromatin structure, sequence motifs, and regulatory elements. These dependencies are also often context-specific and bidirectional, as regulatory sites can affect regions upstream and downstream. BiLSTMs are well-suited for modeling such interactions because they learn dependency structures directly from data without relying on predefined distance-based assumptions. The neural network architecture also supports integrating additional features, such as sequence embeddings, while effectively handling multi-dimensional features and non-linear relationships. By combining forward and backward contexts, BiLSTMs provide a versatile framework for capturing the complex and multidirectional interactions inherent in genomic sequences.

4.2 Model setup

In this section, we present the BiLSTM model to predict DNA methylation patterns on sequencing read $\mathcal{R}_r$. The target sequence Y_r is a binary vector of length T_r, where each element is denoted as $y_t^{(r)} \in \{0, 1\}$ representing the methylation state of site t on sequence Y_r. Thus, this problem is a typical classification problem. Given the sequential nature of the prediction target and input-out characteristics, a many-to-many BiLSTM network is well-suited for this task.

4.2.1 Model structure

The model structure is listed as follows:

1. **Input layer**: The input layer accepts the feature vector $\mathbf{x}_t^{(r)}$ for each site t in Y_r, including methylation level, genomic distance, and other available features such as genomic contexts. It also supports variable-length sequences by padding and masking.

2. **Bidirectional LSTM layer**:

    - Processes input sequences simultaneously in both forward and backward directions to capture dependencies for both directions.

$$\overrightarrow{\mathbf{h}}_t^{(r)} = \text{LSTM}_{\text{forward}}(\mathbf{x}_t^{(r)}, \overrightarrow{\mathbf{h}}_{t-1}^{(r)}), \tag{4.7}$$

$$\overleftarrow{\mathbf{h}}_t^{(r)} = \text{LSTM}_{\text{backward}}(\mathbf{x}_t^{(r)}, \overleftarrow{\mathbf{h}}_{t+1}^{(r)}) \tag{4.8}$$

- The hidden states from forward and backward LSTM layers are concatenated.

$$\mathbf{h}_t^{(r)} = [\overrightarrow{\mathbf{h}}_t^{(r)}; \overleftarrow{\mathbf{h}}_t^{(r)}], \tag{4.9}$$

3. **Fully connected layer**:

   - Maps the concatenated hidden states to output logits and applies a sigmoid activation function to produce methylation probabilities.

$$\mathbf{o}_t^{(r)} = \sigma(\mathbf{W}_{\text{fc}}\mathbf{h}_t^{(r)} + \mathbf{b}_{\text{fc}}), \tag{4.10}$$

   where $\mathbf{W}_{\text{fc}}$ and $\mathbf{b}_{\text{fc}}$ are the weights and biases of the fully connected layer, and σ is the sigmoid activation function. The output $\mathbf{o}_t^{(r)} \in [0, 1]$ represents the predicted probability of methylation.

4.2.2 Loss function design

Our goal for model training is to achieve two objectives simultaneously. First, we aim to capture site-site dependencies, which are implicitly addressed through the use of the LSTM structure. Second, we seek to ensure that the expected predicted states of the read pattern align with the input methylation levels. This alignment is crucial because it addresses the inherent constraint in data summarization, enabling the model to recover the original methylation patterns of individual reads from the summarized site-level methylation levels. To meet these objectives, we propose the following loss function:

1. **Binary Cross-Entropy loss (BCE)**: To ensure that the binary predictions closely match the observed states, we employ the cross-entropy loss as a natural choice for state prediction.

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N_b} \sum_{r=1}^{N_b} \frac{1}{T_r} \sum_{r=1}^{T_r} \left[y_t^{(r)} \log(o_t^{(r)}) + (1 - y_t^{(r)}) \log(1 - o_t^{(r)}) \right], \tag{4.11}$$

where $y_t^{(r)}$ is the true binary methylation state, $o_t^{(r)}$ is the predicted probability, and N_b is the batch size.

2. **Score Alignment loss (Mean Squared Error)**: To preserve the marginal distribution, we aim to ensure that the predicted sequence states maintain the same methylation level for each site as specified in the input features. To achieve this, we define a loss function that encourages alignment between the methylation levels derived from the predictions and those provided in the input.

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N_b} \sum_{r=1}^{N_b} \frac{1}{T_r} \sum_{t=1}^{T_r} \left(\bar{o}_t^{(r)} - m_t^{(r)} \right)^2, \tag{4.12}$$

where $\bar{o}_t^{(r)}$ is the average prediction over multiple stochastic predictions, and $m_t^{(r)}$ is the target methylation score.

3. **Total loss**:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{BCE}} + \beta \mathcal{L}_{\text{MSE}} \tag{4.13}$$

where α and β are weights balancing the two components. By default, we set $\alpha = \beta = 1$. By integrating these two objectives, the model is designed to balance accurate classification and preservation of marginal methylation alignment.

4.2.3 Model training

Data splitting

The dataset was divided into training and testing subsets using an 80/20 split to evaluate the model's generalization capabilities.

Training configuration

The input dimension corresponded to the size of the feature vectors at each time step: 2 for synthetic data and 32 for real WGBS data. The hidden dimension was set to 8 for synthetic data and 32 for real data. Two LSTM layers were stacked to capture complex temporal patterns, and a dropout rate of 0.1 was applied to mitigate overfitting. The output dimension was 1, corresponding to predicting the methylation probability at each genomic site. The model was trained using the Adam optimizer with a learning rate of $\eta = 0.0001$, utilizing mini-batches of size 64.

Training procedure

The model was trained over 50 epochs. In each epoch, the following steps were performed:

1. **Mini-batch processing**: The training data was shuffled and divided into mini-batches. To handle sequences of varying lengths, input sequences were padded to match the length of the longest sequence in each mini-batch, and masks were created to indicate valid positions within sequences for loss calculation. Within each batch, sequences were sorted in descending order of length to improve computational efficiency during training.

2. **Forward pass**: The model processed the input sequences to generate predictions of methylation probabilities at each genomic site.

3. **Loss computation**: The custom loss function, combining Binary Cross-Entropy loss and Mean Squared Error loss, was calculated using the model's predictions and the true labels.

4. **Backward pass and optimization**: Gradients were computed via back-propagation. Gradient clipping was applied with a threshold of 1.0 to prevent exploding gradients, and the model parameters were updated using the Adam optimizer.

5. **Evaluation**: After each epoch, the model's performance was evaluated on the test dataset to monitor training progress and detect potential overfitting.

Training continued until the model converged or until the predefined number of epochs was reached.

4.3 Model prediction and evaluation

4.3.1 Prediction

To create diverse predictions while preserving the desired properties, the model incorporates stochasticity into its outputs. The aim is to ensure that, given the same set of features, the model generates varied predictions that align with marginal expectations and captured dependencies. This is achieved by integrating two techniques: temperature scaling and Gumbel noise.

- **Temperature scaling**: The logits is adjusted by temperature parameter $\tau > 0$ before applying the sigmoid function.

$$\tilde{\mathbf{o}}_t^{(r)} = \frac{\mathbf{o}_t^{(r)}}{\tau}, \tag{4.14}$$

This technique controls the randomness of predictions: A higher τ results in more randomized predictions, while lower values make predictions more deterministic. by default, the τ is set to be 1.

- **Gumbel noise**: Stochastic noise is added to the logits to introduce variability:

$$\mathbf{g}_t^{(r)} = -\ln\left(-\ln\left(\mathbf{u}_t^{(r)}\right)\right), \quad \mathbf{u}_t^{(r)} \sim \text{Uniform}(0,1), \tag{4.15}$$

$$\hat{\mathbf{o}}_t^{(r)} = \tilde{\mathbf{o}}_t^{(r)} + \gamma \mathbf{g}_t^{(r)}, \tag{4.16}$$

where γ is a scaling parameter that determines the noise level. Finally, the probabilities are computed using the sigmoid function:

$$p_t^{(r)} = \text{sigmoid}\left(\hat{\mathbf{o}}_t^{(r)}\right). \tag{4.17}$$

4.3.2 Evaluation

We use the following criteria to evaluate the model training and its ability to preserve both the marginal methylation level for each site and the site-site dependency.

1. **Loss over iterations**: Tracks the total loss ($\mathcal{L}_{\text{total}}$) over training epochs to assess convergence and overfitting in model training.

2. **Marginal methylation level alignment**: Compares the predicted average methylation levels $\bar{o}_t^{(r)}$ with the target methylation level $m_t^{(r)}$.

3. **Site-site dependency preservation**: Computes the autocorrelation of predicted methylation states across sites to evaluate how well the model captures spatial dependencies.

4.4 Simulation study

We conduct simulation studies to evaluate the performance of predictive models under controlled conditions.

4.4.1 Synthetic data generation

Bidirectional generative model

First, we generate binary sequences that contain site-site dependencies with specific marginal probabilities. To achieve this, we utilized a probabilistic generative model inspired by the one-dimensional Ising model [35]. Specifically, this model contains two primary components:

1. **External field** e_t, which biases each site t towards its target marginal probability m_t:

$$e_t = \ln\left(\frac{m_t}{1-m_t}\right) \tag{4.18}$$

2. **Interaction strength** J_t, which introduces dependencies between adjacent sites, with the interaction strength decaying as the distance between adjacent sites d_t increases:

$$J_t = \frac{1}{1+\exp(-(w_0 + w_1 \cdot d_t))} \tag{4.19}$$

   where w_0 controls the base interaction strength and w_1 modulates the decay rate, ensuring closer sites are more likely to influence each other, creating stronger coupling.

Let $\sigma_t \in \{-1, 1\}$ be the hypo-methylation and the hyper-methylation tendency state of site t. The joint probability of observing a state sequence $\sigma = (\sigma_1, \ldots, \sigma_T)$ on a read is defined as

$$\mathbb{P}(\sigma_1, \sigma_2, \ldots, \sigma_T) = \frac{1}{\mathcal{Z}} \exp\left(\sum_{t=1}^{T} e_t\sigma_t + \sum_{t=1}^{T-1} J_{t+1}\sigma_t\sigma_{t+1}\right) \tag{4.20}$$

where $\mathcal{Z}$ is the normalization constant (partition function). In this distribution, the external field e_t anchors each site's probability to its marginal m_t, while the interaction terms J_t modulate correlations between neighboring sites based on distance. By balancing the contributions of the external field and the interaction terms to ensure the former term dominates the probability, the model can effectively capture individual site probabilities and spatial dependencies from both directions (Bidirectional generative model).

Sequence generation using Gibbs sampling

To sample a sequence from the joint distribution, we employed Gibbs sampling to iteratively update the state of each site. Specifically, the Gibbs sampling process calculates the conditional probability of each site given all other sites. For each site t, the total field μ_t acting on it is computed by combining the external field and interactions with neighboring sites:

$$\mu_t = e_t + J_t \cdot \sigma_{t-1} \cdot \mathrm{I}(t > 1) + J_{t+1} \cdot \sigma_{t+1} \cdot \mathrm{I}(t < T), \tag{4.21}$$

where $\mathrm{I}(\cdot)$ is an indicator function ensuring valid boundary conditions. The state σ_t is then updated by

$$\sigma_t = \begin{cases} 1 \text{ with probability } \frac{1}{1+\exp(-\mu_t)} \\ -1 \text{ otherwise} \end{cases} \tag{4.22}$$

By iteratively updating each site over multiple iterations, the sequence converges to a stationary distribution that balances external field and interaction terms. This approach offers a controlled and flexible framework for generating synthetic binary data that preserves the specified marginal probabilities while reflecting spatial dependencies between sites. The full sequence generation procedure is detailed in Algorithm 5.

Algorithm 5 Bidirectional generative model via Gibbs Sampling

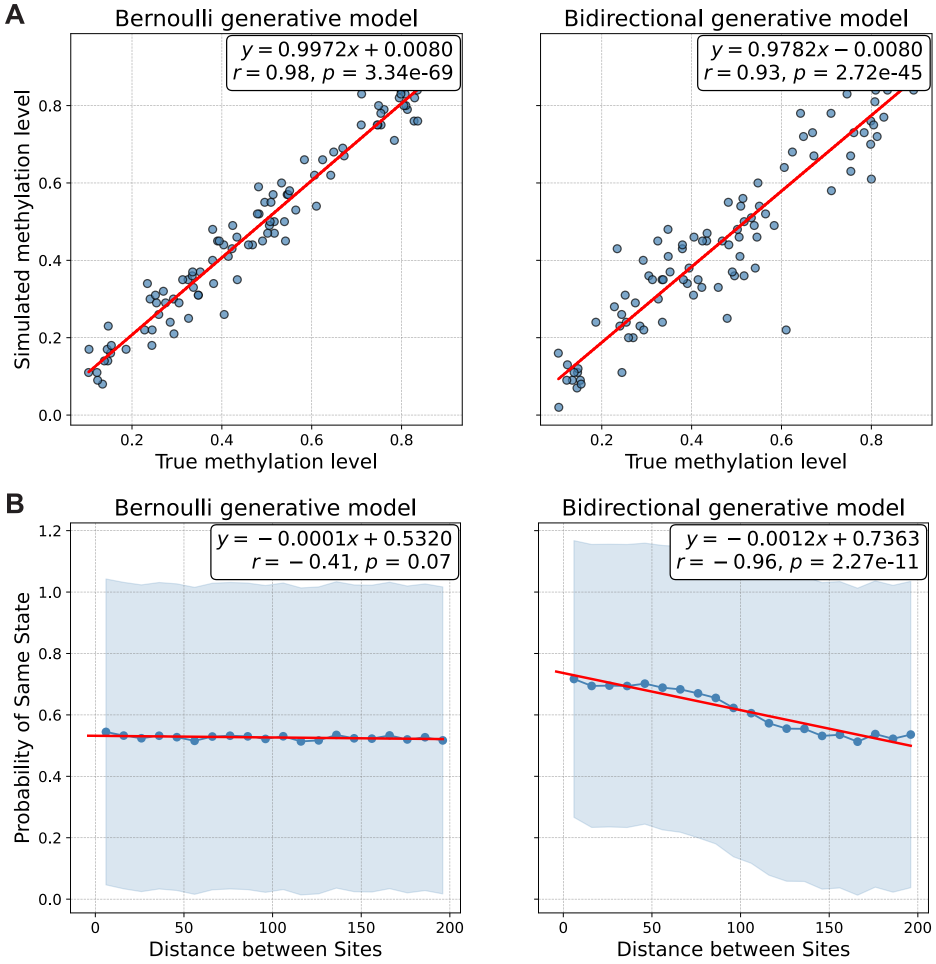
```
Require:
  $M = (m_1, \dots, m_T) \in [0,1]^T$: Marginal probabilities for $T$ sites;
  $D = (d_1, \dots, d_T) \in \mathbb{N}^T$: Distances of adjacent sites, with $d_1$ set to be 0;
  $n\_iter$: Number of Gibbs sampling iterations;
  $w_0, w_1$: Interaction strength parameters.
Ensure: : the generated binary sequence $Y = (y_1, \dots, y_T) \in \{0,1\}^T$
  Compute external field: $e_t = \log\left(\frac{m_t}{1-m_t}\right)$ for $t = 1, \dots, T$
  Compute interaction strengths: $J_t = \frac{1}{1+\exp(-(w_0+w_1 \cdot d_t))}$ for $t = 1, \dots, T$
  Initialize $\sigma = (\sigma_1, \dots, \sigma_T) \in \{-1,1\}^T$ randomly
  for iter = $1, \dots, n\_iter$ do
      for $t = 1, \dots, T$ do
          Compute total field:
              $$\mu_t = e_t + J_t \cdot \sigma_{t-1} \cdot \mathrm{I}(t > 1) + J_{t+1} \cdot \sigma_{t+1} \cdot \mathrm{I}(t < T)$$
          Update state:
              $$\sigma_t = \begin{cases} 1 & \text{with probability } \frac{1}{1+\exp(-\mu_t)}, \\ -1 & \text{otherwise.} \end{cases}$$
      end for
  end for
  Convert $\sigma$ to $Y$: $y_t = \frac{\sigma_t+1}{2}$ for all $t = 1, \dots, T$
  return $Y$
```

Simulation validation

To validate that the synthetic data has matched marginal probability and contains spatial dependencies, we generated 100 synthetic sequences, each consisting of 100 sites. Distances between adjacent sites were sampled from a uniform distribution between 1 and 200, while methylation levels were drawn from a uniform distribution between 0.1 and 0.9. For the bidirectional generative model, the interaction strength weights were set to $w_0 = 5$ and $w_1 = -0.05$, and the number of Gibbs sampling iterations (n_iter) was set to 100. We also simulated sequences using the methylation level based on the independent Bernoulli model for comparison. Both models were evaluated for their ability to preserve the marginal distribution (methylation level estimates) and capture site-site dependencies in the synthetic sequences.

Figure 4.2 demonstrates the key differences between these two data generation models. Panel A shows both models accurately preserve the marginal methylation levels, with high correlations between prediction average and true levels ($r = 0.98$ for the Bernoulli model and $r = 0.93$ for the bidirectional model). Panel B evaluates site-site dependencies, showing the probability of shared states as a function of distance by binning the distances. The Bernoulli model exhibits no significant dependency ($r = -0.41$, $p = 0.07$), while the bidirectional model captures a clear, negative correlation ($r = -0.96$, $p = 2.27 \times 10^{-11}$). These results validate the bidirectional generative model's ability to generate sequences with spatial dependencies with matched marginal distributions.

Figure 4.2: Comparison of two synthetic data generation models. Both models can generate sequences that have average methylation states match the marginal distribution, however, only bidirectional generative model can introduce spatial dependency into the synthetic data.

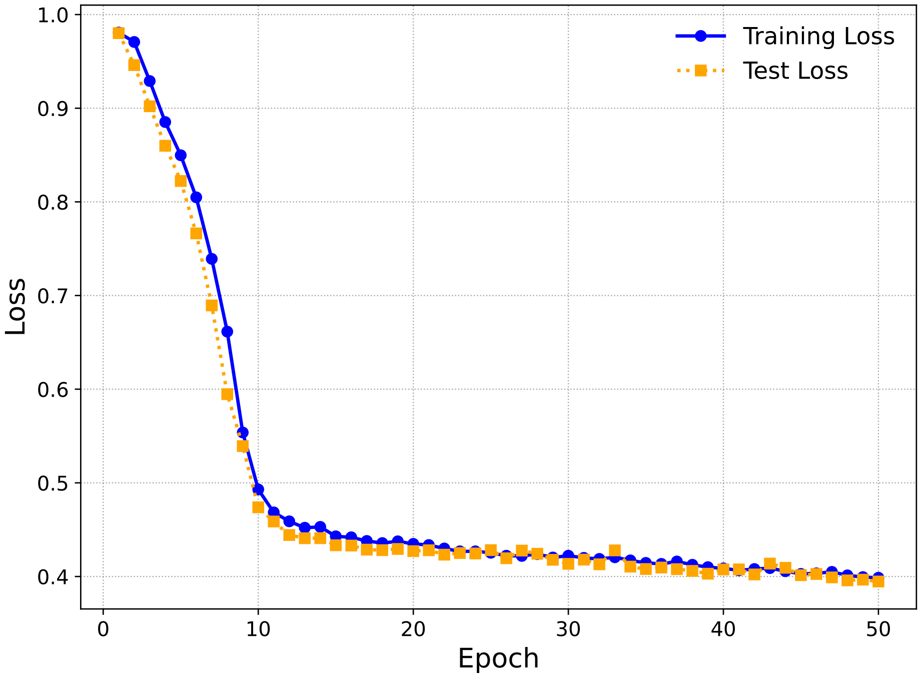


4.4.2 Model performance evaluation

Loss curve over iteration

We trained and evaluated the BiLSTM model using a synthetic dataset of 10,000 sequences generated by the bidirectional generative model, with methylation levels sampled in a manner consistent with chapter 3. Of these, 8,000 sequences were used for training, and 2,000 were reserved for testing. The training and testing loss curves, shown in Figure 4.3, indicate a smooth convergence over 50 epochs. Both the training and test losses decrease rapidly in the initial epochs, followed by a gradual plateau and stabilization of around 0.4, suggesting that the model successfully learns the underlying data patterns. The close alignment between the training and test loss further suggests good model generalization.

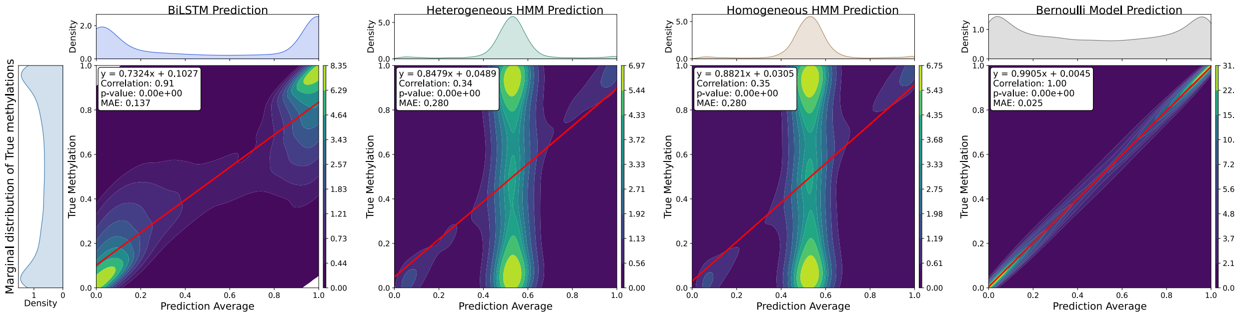
Figure 4.3: Training and test loss convergence of the BiLSTM model on synthetic data.



Marginal distribution preservation

To demonstrate that the BiLSTM effectively preserves the marginal distribution, we compared the average predicted methylation states with the true methylation levels (Figure 4.4). The BiLSTM achieved a strong correlation ($r = 0.91$) and a low mean absolute error (MAE $= 0.137$). For comparison, we evaluated other models. The heterogeneous HMM ($r = 0.34$, MAE $= 0.280$) and homogeneous HMM ($r = 0.35$, MAE $= 0.280$) performed significantly worse in capturing the marginal distribution. Importantly, these models failed to produce a bi-modal distribution of marginal methylation levels, as shown in the true methylation levels. This limitation arises because they rely solely on the initial methylation state in their models, which restricts their ability to capture the full distribution. While the Bernoulli model achieved the best correlation ($r = 1.00$) and the lowest error (MAE $= 0.025$), it does not model dependencies between sites, which will be displayed in the next section.

Figure 4.4: Evaluation of marginal distribution preservation across models on synthetic data.

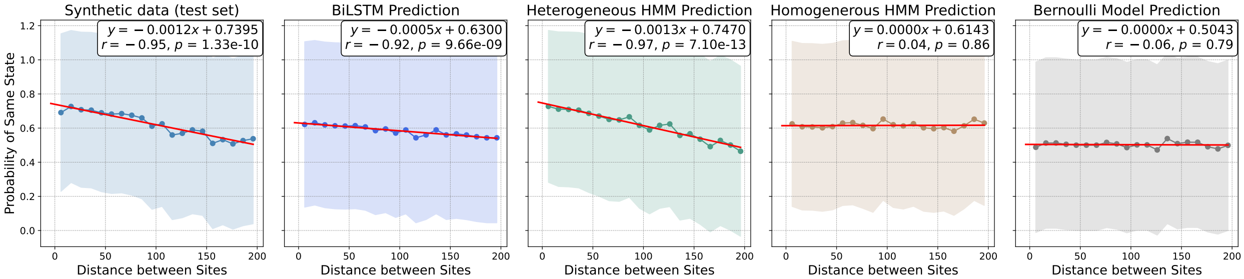


Site-site dependency capture

The ability to capture site-site dependency was evaluated by examining the probability of adjacent sites sharing the same state as a function of distance, analyzed within binned windows. The BiLSTM demonstrated strong performance, achieving a correlation of $r = -0.92$, closely aligning with the synthetic data's true correlation ($r = -0.95$), thereby effectively modeling spatial dependencies in methylation levels.

Among the comparison models, the heterogeneous HMM performed comparably well ($r = -0.97$), reflecting its ability to account for distance-dependent transitions. In contrast, the homogeneous HMM ($r = 0.04$) and the Bernoulli model ($r = -0.06$) failed to capture meaningful dependencies. This lack of correlation arises from two key factors: (1) methylation levels within the window bins are not intrinsically correlated, and (2) neither model successfully captures the relationship between distance and site similarity. Specifically, the homogeneous HMM assumes uniform transition probabilities irrespective of distance, while the Bernoulli model treats sites as entirely independent, disregarding spatial relationships. As a result, neither model reproduces the correlation patterns inherent to the data. Notably, the Bernoulli model also exhibited the lowest probability of adjacent sites sharing the same state. This outcome stems from its independence assumption, which underestimates correlations and leads to a more dispersed state distribution.

Figure 4.5: Evaluation of site-site dependency capturing across models on synthetic data.



Collectively, the simulation demonstrates that the BiLSTM model strikes a balance between marginal distribution preservation and site-site dependency capturing, outperforming HMM-based approaches in preserving the marginal distribution and competing closely with the heterogeneous HMM in capturing spatial dependencies. Although the Bernoulli model achieves near-perfect marginal accuracy, its inability to incorporate dependencies limits its suitability for data with spatial relationships.

4.5 Application to WGBS data

To extend the model evaluation on the real whole-genome bisulfite sequencing (WGBS) data, we applied it to predict methylation patterns at individual CpG sites. Each site was represented by a 32-dimensional feature vector designed to comprehensively incorporate methylation, spatial, and sequence-context information. The features included:

- Methylation Level: Summarized methylation status of each CpG site based on sequencing reads.

- Genomic Distance: The number of base pairs between consecutive CpG sites, providing spatial context to account for physical proximity and its influence on site-site dependency.

- Nucleotide Identity: The base at the current position, encoded as 0 for all cytosines (the focus of our analysis).

- Sequence Context: The cytosine context (`CG`, `CHG`, or `CHH`), capturing sequence-level variations known to influence methylation patterns.

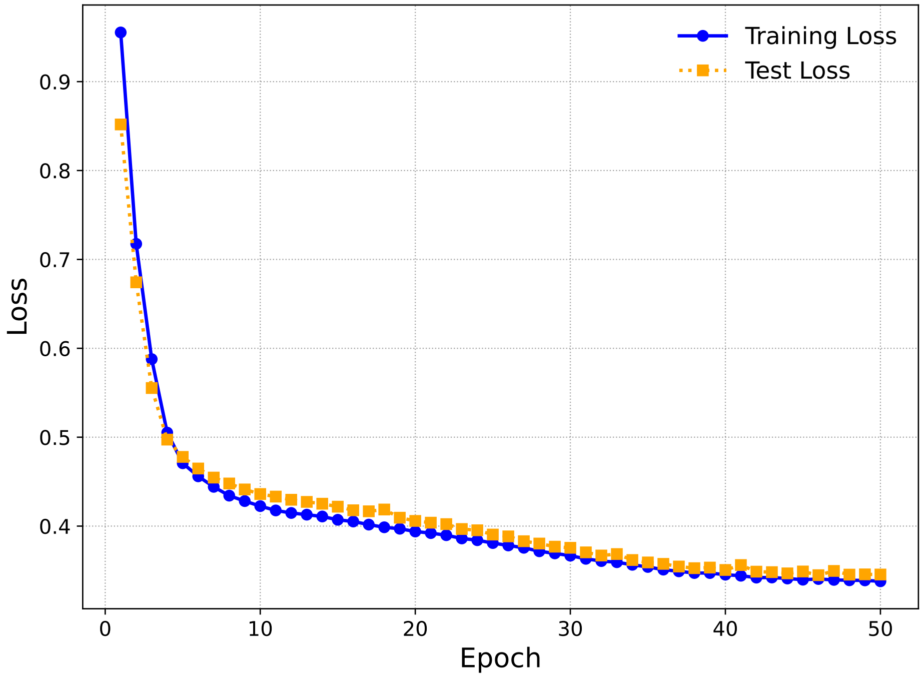
- Flanking Sequence: One-hot encoded representations of the three nucleotides upstream and downstream (forming a 7-mer) to represent the local genomic environment.

This feature set effectively integrates key biological and spatial characteristics to model dependencies in methylation patterns.

Loss curve over iteration

Following model training, the loss curve for the BiLSTM model trained on WGBS data demonstrates smooth convergence over 50 epochs (Figure 4.6). Both training and test losses decrease rapidly during the initial epochs and stabilize in the following epochs, reflecting the model's effective learning of WGBS data patterns. The close alignment between training and test losses further underscores the model's strong generalization to unseen data, indicating minimal overfitting.

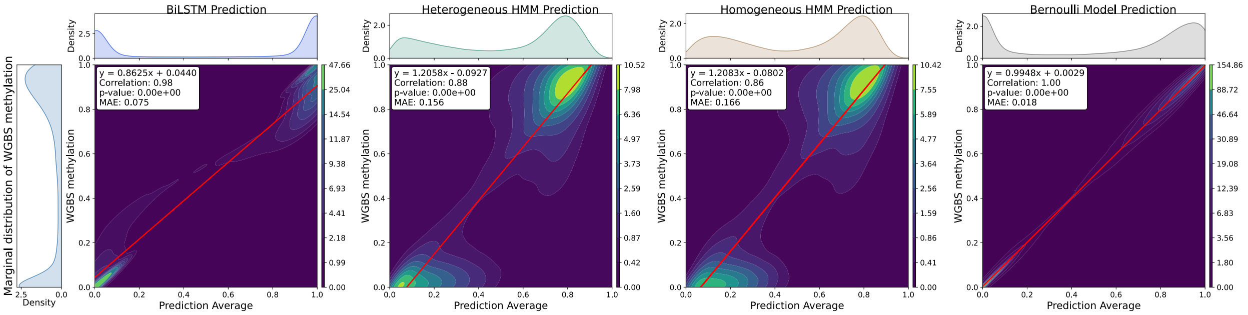
Figure 4.6: Training and test loss convergence of the BiLSTM model on WGBS data.



Marginal distribution preservation

The preservation of the marginal methylation distribution in WGBS data was evaluated by comparing predicted methylation levels to true methylation levels across models (Figure 4.7). The BiLSTM achieved a high correlation ($r = 0.98$) and a low mean absolute error (MAE $= 0.075$), demonstrating its ability to balance accuracy with dependency modeling. In comparison, the heterogeneous HMM ($r = 0.88$, MAE $= 0.156$) and homogeneous HMM ($r = 0.86$, MAE $= 0.166$) exhibited lower performance in capturing the marginal distribution, likely due to their limited utilization of methylation levels in the modeling process. Lastly, the Bernoulli model achieved perfect correlation ($r = 1.00$) and the lowest MAE (0.018), reflecting its precise preservation of marginal probabilities, albeit without accounting for dependencies.

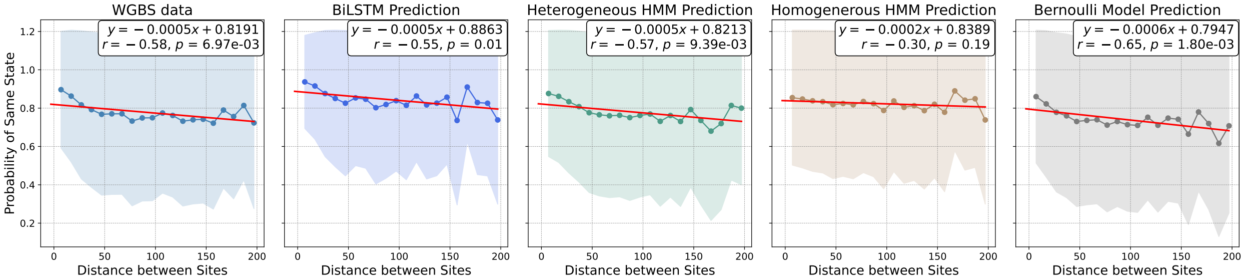
Figure 4.7: Evaluation of marginal distribution preservation across models on WGBS data.



Site-site dependency capture

Site-site dependencies were evaluated by examining the probability of adjacent sites sharing the same state as a function of distance (Figure 4.8). The synthetic WGBS data exhibits a strong negative correlation ($r = -0.58$), which the BiLSTM closely replicates ($r = -0.55$), demonstrating its ability to capture spatial dependencies. Similarly, the heterogeneous HMM performs well in modeling this relationship ($r = -0.57$). In contrast, the homogeneous HMM ($r = -0.30$) fails to accurately capture this dependency. The Bernoulli model, with a correlation of -0.65, likely reflects the coupling of methylation levels within the window bin. However, its correlation and regression slope deviate from those of the real WGBS data, indicating that it does not intrinsically capture the site-site dependency. Collectively, these results highlight the BiLSTM's effectiveness in modeling spatial dependencies while preserving marginal accuracy on the real WGBS data.

Figure 4.8: Evaluation of site-site dependency capturing across models on WGBS data.



Chapter 5

Conclusion

5.1 Summary

DNA methylation patterns exhibit complex dependencies between adjacent sites, influenced by factors such as genomic distance, chromatin structure, and enzymatic activity. Conventional bisulfite sequencing data analysis often ignores these site-site dependencies, limiting their potential in mechanistic studies and clinical applications. This thesis explored two complementary methodologies for modeling the site-site dependencies: a statistical modeling approach using heterogeneous Hidden Markov Models (HMM), and a machine learning approach employing Bidirectional Long Short-Term Memory (BiLSTM) networks.

In chapter 3, we introduced the heterogeneous HMM, an extension of the classical homogeneous HMM framework that incorporates genomic distance into transition probabilities. This approach reflects the biological intuition that adjacent CpG sites exhibit stronger dependencies due to their similar thermodynamic microenvironment. By explicitly modeling the dependency structure as a function of the genomic distance between adjacent sites, we derived a parameter estimation procedure utilizing the Expectation-Maximization (EM) algorithm. Using synthetic data with site-site dependencies, we demonstrated that heterogeneous HMM has superior performance in model fitting, parameter estimation accuracy, and capturing distance-related dependency patterns compared to homogeneous HMM, at a cost of higher computational demand. When applied to real WGBS data, the heterogeneous HMM also outperformed the homogeneous HMM in both data fitting and dependency preservation, effectively capturing the diminishing dependency as genomic distance increases.

This model thus provides a more accurate representation of methylation patterns than its homogeneous counterpart.

While effective in modeling distance-dependent methylation patterns, the heterogeneous HMM's reliance on a simplified parametric dependency structure and first-order Markovian assumption limits its ability to capture complex and long-range dependencies. To address these challenges, chapter 4 introduces a deep learning approach using the Bidirectional Long Short-Term Memory (BiLSTM) network. This model leverages recurrent neural network architecture to implicitly learn sequential dependencies from methylation data, considering both forward and backward directions. By incorporating a rich set of features—including methylation levels, genomic distances, and other available features such as sequence context embeddings—the BiLSTM accurately matches marginal methylation probabilities while effectively preserving site-site dependencies. Simulation studies and applications to WGBS data demonstrate that the BiLSTM outperforms both homogeneous and heterogeneous HMMs in aligning with marginal methylation levels. While the independent Bernoulli model excels in preserving marginal distributions, the BiLSTM's strength lies in its ability to capture intricate site-site dependencies. This deep learning approach thus provides a powerful and flexible alternative, extending beyond the constraints of parametric models.

Together, these complementary approaches underscore the potential of employing statistical modeling and machine learning to tackle the complexities of DNA methylation data. The heterogeneous HMM offers interpretability and precision for modeling distance-related dependencies. On the other hand, the BiLSTM network introduces the flexibility and versatility to integrate diverse features and capture intricate patterns. Both approaches can advance our ability to characterize the methylation pattern dynamics observed in real-world data.

5.2 Future directions

Despite these advancements, several improvements and extensions can be achieved to enhance the models further and broaden their applicability.

1. Extending the heterogeneous HMM From a modeling perspective, the current heterogeneous HMM assumes that the entire methylation pattern is determined solely by the initial probability (at the first site) and the genomic distance between sites. This assumption can be extended to incorporate the methylation levels of all sites as covariates, enabling a more nuanced understanding of site-site interactions and enhancing the model's ability to capture true dependencies arising from genomic distances. Additionally, recent studies have demonstrated the utility of prior information in enhancing the modeling of complex dependencies [36]. For the heterogeneous HMM, priors derived from external datasets—such as pre-estimated distributions of transition probabilities or emission parameters— can regularize parameter estimation. This approach may facilitate faster convergence and yield more robust results, particularly in sparse or noisy data scenarios.

On the computational side, scalability remains a significant challenge when handling large datasets, such as the full WGBS data with hundreds of millions of observations. Distributed computing can address this challenge by parallelizing workloads across multiple processors or nodes, reducing runtime and improving scalability. Additionally, frameworks like Auto-Encoding Variational Bayes (AEVB)[37] offer a solution by leveraging stochastic optimization and modern GPU hardware acceleration to enable faster and more efficient inference [38] . Together, these strategies can enhance computational efficiency and make the model more practical for high-throughput applications.

2. Enhancing Neural Network frameworks The BiLSTM framework focuses on predicting sequences that align with observed marginal distributions while introducing site-site dependencies. While this thesis did not focus on exploring the factors contributing to methylation concordance, future work could incorporate explainable machine learning techniques to identify key features driving such dependencies. This approach would provide valuable insights into the biological mechanisms underlying DNA methylation, enhancing our understanding of the factors that regulate its patterns and dynamics.

Additionally, we recognize the potential of other advanced generative models for site-site dependency modeling. Transformer architectures [39], with their attention mechanisms, are well-suited for capturing complex dependencies across sites. Diffusion models [40], renowned for generating realistic data, present another promising avenue for simulating realistic methylation patterns while preserving marginal distributions and site-site dependencies. Exploring these frameworks could significantly enhance the scope and precision of methylation pattern analysis, paving the way for more advanced applications and discoveries.

3. Applications in genomic and epigenomic analysis Future work should also focus on extending these models to broader applications in genomic and epigenomic research. The presence of site-site dependencies underscores the potential to borrow information across sites, enabling tasks such as missing value imputation in sparse datasets, including single-cell methylation sequencing, where leveraging local dependencies can improve data quality and enhance downstream analyses. Additionally, incorporating dependency structures could refine the detection of differentially methylated regions (DMRs), increasing statistical power and precision. However, to ensure the reliability of these applications, it is critical to prioritize extending the models to more diverse datasets to validate their generalizability. Without thorough testing, biases from the training data could propagate during tasks like imputation, compromising the integrity of results. Addressing these challenges will expand the practical applications of these models and strengthen the statistical rigor in epigenomic analyses.

In summary, this thesis introduced two novel methodologies for modeling site-site dependencies in DNA methylation data using statistical and machine-learning approaches. The heterogeneous HMM provides an interpretable framework for capturing distance-dependent dependencies, while the BiLSTM offers flexibility to integrate diverse features and model complex patterns. Together, these methods enhance our ability to characterize methylation dynamics and support practical applications in epigenomic research, such as realistic synthetic data generation and missing value imputation. Future work can build on these approaches by addressing current limitations and extending their application to new domains, paving the way for deeper biological insights and translational advances in epigenomics.

References

[1] Suhua Feng et al. "Conservation and divergence of methylation patterning in plants and animals". In: *Proceedings of the National Academy of Sciences* 107.19 (2010), pp. 8689–8694.

[2] Alexandra L Mattei, Nina Bailly, and Alexander Meissner. "DNA methylation: a historical perspective". In: *Trends in Genetics* 38.7 (2022), pp. 676–707.

[3] Maxim VC Greenberg and Deborah Bourc'his. "The diverse roles of DNA methylation in mammalian development and disease". In: *Nature reviews Molecular cell biology* 20.10 (2019), pp. 590–607.

[4] Lisa D Moore, Thuc Le, and Guoping Fan. "DNA methylation and its basic function". In: *Neuropsychopharmacology* 38.1 (2013), pp. 23–38.

[5] Naomi Moris, Cristina Pina, and Alfonso Martinez Arias. "Transition states and cell fate decisions in epigenetic landscapes". In: *Nature Reviews Genetics* 17.11 (2016), pp. 693–703.

[6] Zelin Jin and Yun Liu. "DNA methylation in human diseases". In: *Genes & diseases* 5.1 (2018), pp. 1–8.

[7] Wenbin Guo et al. "Type-2 diabetes biomarker discovery and risk assessment through saliva DNA methylome". In: *medRxiv* (2024).

[8] Samareh Younesian et al. "The DNA methylation in neurological diseases". In: *Cells* 11.21 (2022), p. 3439.

[9] Shawn J Cokus et al. "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning". In: *Nature* 452.7184 (2008), pp. 215–219.

[10] Mary E Sehl et al. “Systematic dissection of epigenetic age acceleration in normal breast tissue reveals its link to estrogen signaling and cancer risk”. In: *bioRxiv* (2024), pp. 2024–10.

[11] Xiaoke Hao et al. “DNA methylation markers for diagnosis and prognosis of common cancers”. In: *Proceedings of the National Academy of Sciences* 114.28 (2017), pp. 7414–7419.

[12] Ornella Affinito et al. “Nucleotide distance influences co-methylation between nearby CpG sites”. In: *Genomics* 112.1 (2020), pp. 144–150.

[13] C Anthony Scott et al. “Identification of cell type-specific methylation signals in bulk whole genome bisulfite sequencing data”. In: *Genome biology* 21 (2020), pp. 1–23.

[14] Babraham Bioinformatics. *Sherman*. https://github.com/FelixKrueger/Sherman/.

[15] Qing Xie et al. “A Bayesian framework to identify methylcytosines from high-throughput bisulfite sequencing data”. In: *PLoS Computational Biology* 10.9 (2014), e1003853.

[16] Giulia Piaggeschi et al. “MethylFASTQ: a tool simulating bisulfite sequencing data”. In: *2019 27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*. IEEE. 2019, pp. 334–339.

[17] Colin Farrell et al. “BiSulfite Bolt: A bisulfite sequencing analysis platform”. In: *GigaScience* 10.5 (2021), giab033.

[18] Owen JL Rackham et al. “WGBSSuite: simulating whole-genome bisulphite sequencing data and benchmarking differential DNA methylation analysis tools”. In: *Bioinformatics* 31.14 (2015), pp. 2371–2373.

[19] Ren-Hua Chung and Chen-Yu Kang. “pWGBSSimla: a profile-based whole-genome bisulfite sequencing data simulator incorporating methylation QTLs, allele-specific methylations and differentially methylated regions”. In: *Bioinformatics* 36.3 (2020), pp. 660–665.

[20] Jianfeng Xu et al. "Cellular Heterogeneity–Adjusted cLonal Methylation (CHALM) improves prediction of gene expression". In: *Nature communications* 12.1 (2021), p. 400.

[21] Olga Chervova et al. "The Personal Genome Project-UK, an open access resource of human multi-omics data". In: *Scientific data* 6.1 (2019), p. 257.

[22] Shifu Chen et al. "fastp: an ultra-fast all-in-one FASTQ preprocessor". In: *Bioinformatics* 34.17 (2018), pp. i884–i890.

[23] Marta Byrska-Bishop et al. "High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios". In: *Cell* 185.18 (2022), pp. 3426–3440.

[24] R Durbin. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. 1998.

[25] Mark Borodovsky and James McIninch. "GENMARK: parallel gene recognition for both DNA strands". In: *Computers & chemistry* 17.2 (1993), pp. 123–133.

[26] Jason Ernst and Manolis Kellis. "ChromHMM: automating chromatin-state discovery and characterization". In: *Nature methods* 9.3 (2012), pp. 215–216.

[27] Timothy L Bailey et al. "MEME SUITE: tools for motif discovery and searching". In: *Nucleic acids research* 37.suppl_2 (2009), W202–W208.

[28] Michael Scherer et al. "Quantitative comparison of within-sample heterogeneity scores for DNA methylation data". In: *Nucleic acids research* 48.8 (2020), e46–e46.

[29] John C Marioni, Natalie P Thorne, and Simon Tavaré. "BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data". In: *Bioinformatics* 22.9 (2006), pp. 1144–1146.

[30] Diane P Genereux et al. "Errors in the bisulfite conversion of DNA: modulating inappropriate-and failed-conversion frequencies". In: *Nucleic acids research* 36.22 (2008), e150–e150.

[31] Nicholas Stoler and Anton Nekrutenko. “Sequencing error profiles of Illumina sequencing instruments”. In: *NAR genomics and bioinformatics* 3.1 (2021), lqab019.

[32] Sae Rom Hong and Kyoung-Jin Shin. “Bisulfite-converted DNA quantity evaluation: a multiplex quantitative real-time PCR system for evaluation of bisulfite conversion”. In: *Frontiers in genetics* 12 (2021), p. 618955.

[33] Mengchi Wang et al. “Identification of DNA motifs that regulate DNA methylation”. In: *Nucleic Acids Research* 47.13 (2019), pp. 6753–6768.

[34] S Hochreiter. “Long Short-term Memory”. In: *Neural Computation MIT-Press* (1997).

[35] Roy J Glauber. “Time-dependent statistics of the Ising model”. In: *Journal of mathematical physics* 4.2 (1963), pp. 294–307.

[36] Ziyi Song et al. “Clustering computer mouse tracking data with informed hierarchical shrinkage partition priors”. In: *Biometrics* 80.4 (2024), ujae124.

[37] Diederik P Kingma. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[38] Yiwei Gong, Susanna B Mierau, and Sinead A Williamson. “Detecting State Changes in Functional Neuronal Connectivity using Factorial Switching Linear Dynamical Systems”. In: *arXiv preprint arXiv:2411.04229* (2024).

[39] A Vaswani. “Attention is all you need”. In: *Advances in Neural Information Processing Systems* (2017).

[40] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.