Saarland University
Center for Bioinformatics
Master's Program in Bioinformatics

Master's Thesis in Bioinformatics

Design and calibration of stochastic models for DNA methylation patterns

submitted by

Andrea Kupitz

on March 2019

Supervisor
Prof. Dr. Verena Wolf

Advisor
M.Sc. Alexander Lück

Reviewers
Prof. Dr. Verena Wolf
Prof. Dr. Volkhard Helms

Zentrum für Bioinformatik

Kupitz, Andrea
Design and calibration of stochastic models for DNA methylation patterns
Master's Thesis in Bioinformatics
Universität des Saarlandes
Saarbrücken, Germany
March 2019

Declaration

I hereby confirm that this thesis is my own work and that I have documented all sources used.

I hereby declare that the submitted digital and hardcopy versions of this thesis correspond to each other. I give permission to the Universität des Saarlandes to duplicate and publish this work.

Saarbrücken, on March 2019

Andrea Kupitz

Abstract

The expression of genes in the human genome is not only based on the DNA sequence; it relies on epigenetic modifications like DNA-methylation. Hereby, the gene expression is inactivated by the binding of methyl-groups to the cytosine of a cytosine-phosphate-guanine (CpG) dinucleotide at the promoter region of the concerned gene. The binding is performed by specific enzymes - the DNA methyltransferases (DNMTs).

In this thesis, we model the methylation of DNA using an Markov Chain Monte Carlo (MCMC) algorithm and estimate the corresponding parameters with Maximum Likelihood Estimation (MLE). Alternatively, parameter estimation is performed with an implementation of the approximative Bayesian computation (ABC) method. Moreover, different distance functions are provided to compare distributions of methylation patterns.
We find differences in the function of the different DNMTs that are consistent with current biological findings. Using this model, some parameters are difficult to identify because they seem to be conditionally dependant.

Acknowledgments

I would like to thank deeply Prof. Dr. Verena Wolf and Alexander Lück, who introduced me to my master theme and always offered me their advice and support which I appreciated very much.

I owe thanks to both of them for the opportunity to develop this master thesis at the *Center for Bioinformatics*.

Furthermore I would like to thank all other people who put in touch with me during the development of this master thesis and who were very helpful in providing information.

Contents

List of Tables

List of Figures

List of Abbreviations

5mC 5-methylcytosine . . . 1
5hmC 5-hydroxymethylcytosine . . . 4
ABC approximative Bayesian computation . . . ii
BS-seq bisulfite-sequencing . . . 11
bp basepair . . . 12
CGI CpG island . . . 1
CpG cytosine-phosphate-guanine . . . ii
CTMC continuous-time Markov Chain . . . 6
DNMT DNA methyltransferase . . . ii
DTMC discrete-time Markov Chain . . . 6
GpC guanine-phosphate-cytosine . . . 3
HMM Hidden Markov Model . . . 7
KO knock-out . . . 13
MCMC Markov Chain Monte Carlo . . . ii
MGMT O6-methylguanine-DNA methyltransferase . . . 13
MLE Maximum Likelihood Estimation . . . ii
PCR polymerase chain reaction . . . 4
RV random variable . . . 5
WT wild-type . . . 14

Chapter 1

Introduction

When the expression of genes is studied, DNA methylation is a major issue which should be considered. This DNA modification mainly occurs at the 5 position of the pyrimidine ring of a cytosine base in the DNA[1]. The methylated cytosine is thus often called 5-methylcytosine (5mC). Whereas 5mC is rather rare over the whole genome, there exist small regions with high methylation frequency, so called CpG islands (CGIs). Variations in the methylation pattern of CGIs are related to changes in gene expression and cancer[2]. Additionally, methylcytosine is associated to the process of genomic imprinting and X-chromosome inactivation[3].
To study the transition of methylations from one cell generation to another, one needs to focus on DNA methyltransferases (DNMTs). The general conformation is that there exist two kinds of methyltransferases, which may be discriminated by their function. DNMT1 seems to copy methylations from the parent to the daughter strand and its activity is related to the replication machinery. While DNMT3 may work at different positions in the DNA without any environmental conditions[1]. In order to deepen the knowledge of DNA methylations and their spreading, computer methods are used to simulate their behaviour. More precisely, the methylation rates and their dependence on other system parameters are estimated. Therefore, it is made use of the fact that there is knowledge about the methylation state of specific loci from biological experiments.

In [4] developed one of the first models for DNMT simulation. In [5] and [6] the general idea is extended, allowing the possibility of errors in the data. 2016 in [7] a model was published that includes demethylation events during the replication process. But all these models do not include any location-dependency of the methylation rates despite the data allows the assumption.[5]
This idea of neighbourhood- and location-dependency was regarded in multiple approaches. In [8], paper from 2012, the processivity of DNMTs was investigated, availing the methylation states of all preceding positions and trying to infer the binding state of the enzymes. However, the approach fails to identify all parameters.
Integrating the limitations of the previous method, Bonello et al.[9] compare different location- and neighbour-dependent models to spot the model which seems to fit best to the real-world data. They conclude, that a model which respects the neighbouring positions is most likely to result in the desired data.

Recently, a aimilar approach[10] was published that takes direct neighbouring methylation states into account. The realized experiments reason a dependence on the left-, but not on the rightmost neighbour. Nevertheless, the binding state of the DNMTs is not kept track of and thus the association of the enzyme to the replication process is still unknown.

Here we will present an approach which is based on the same parameters for DNMTs as in the paper by Fu et al.(2012)[8]. The related works use different models and methods to estimate the process parameters. We simulate the transmission of methylations in one way, but compare the computational methods for parameter estimation. We show that we are able to infer the function of the two kinds of methyltransferases and that their their methylation rates are differentiable. These results agree to common, recent findings and suppositions, which we are able to prove.
TODO: extend... (why is this model useful)

Chapter 2

Background

2.1 DNA methylation

Epigenetic studies the heritable information not relying on changes in the DNA sequence and influencing the organisms phenotype. There exist two kinds of epigenetic modifications: Chromatin and DNA modifications. Chromatin is the three-dimensional arrangement of DNA and a histone protein. The modification of Chromatin is performed by binding of an amino group or RNA[11].

Figure 2.1: *Methylation states; each figure shows a CpG/CpG-dyad on the parental and the daughter strand; each circle represents one CpG; plane red circles are methylated CpGs; (a) unmehtylated, (b) hemimethylated (upper strand unmethylated, lower strand methylated), (c) hemimethylated (vice versa), (d) fully methylated CpGs.*

The most common DNA modification is the addition of one methyl group to the fifth position in the cytosine ring of DNA. Methylation occurs mainly at cytosine-phosphate-guanines (CpGs), where a cytosine nucleotide (C) is followed by a guanine nucleotide (G) in the DNA sequence. By the complementary basepairing of a CpG to a guanine-phosphate-cytosine (GpC) on the other strand, there result two possible methylation positions at this CpG/CpG dyad. Hence, four methylation states may be discriminated. Either the dyad is unmethylated, fully methylated or one of the two cytosines is methylated and thus the dyad is hemimethylated. Figure 2.1 shows the possible methylation states. By the sequence of methylations over the whole genome, an individual methylation pattern can be retrieved for each organism[1].

Whereas the majority of CpGs in mammals are methylated, so called CpG islands (CGIs) are rather unmethylated. A CGI is defined as region in the genome, where the frequency of CpGs is very high. Figure **??** shows an example. The CGIs are associated to gene regulation as they are often located at

CCGAACGTCGCGCGTCA
GGCTTGCAGCGCGCAGT

Figure 2.2: *CpG-island with increased cytosine and guanine frequency; CpGs are marked in red.*

the promoter region of genes. Hereby, cytosine methylation inhibits the gene expression; leading to changes in the methylation pattern of the DNA are effecting diseases like cancer[2]. Hypomethylation of repepetetive elements for example results in an unstable DNA and may increase the risk of cancer; as also the hypermethylation of cancer suppressor genes.[1] Finally, DNA methylation is responsible for genomic imprinting and X-chromosome inactivation, whereby one of the two alleles respectively is transcribed and the other inactivated by methylation. Dysregulation may contribute to diseases like Prader-Willi syndrome, Angelman syndrome and cancer[3].

To detect methylcytosine, the most popular method is bisulfite sequencing, incorporated with polymerase chain reaction (PCR). During bisulfite treatment, unmethylated cytosines are transformed to uracil, whereas methylated ones stay unaffected. PCR sequencing is used to retrieve the base sequence[12]. Nevertheless, bisulfite sequencing has multiple drawbacks. First, it depends on the correct conversion of cytosines to uracil, which is not given as errors during the bisulfite treatment may occur. Second, a differentiation of 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) is not possible because the two molecules respond in the same way to bisulfite treatment[13]. 5hmC is another DNA modification, which is associated to the process of active and passive demethylation[14][7]. See section 2.6 for further details.

The transfer of the methyl group to the DNA is performed by DNA methyltransferases (DNMTs). Five different DNMTs are distinguished in mammals: DNMT1, DNMT2, DNMT3A, DNMT3B and DNMT3L[1].
DNMT1 is commonly associated to maintanance methylation and the DNA replication process. Thereby, not only the DNA of the cell is transmitted from one cell generation to another, but also the methylation patterns. It is supposed, that DNMT1 has a preference for hemimythelated DNA. Subsequently, DNMT1 transfers the methylation by methylating the positions on the daughter strand that are methylated at the same position on the parental strand[1].
DNMT2 is negligible if human DNA methylation is considered, because it methylates small RNA at the anticodon loop[15].
DNMT3A and DNMT3B are de novo methyltransferases, which work during the early embryonic development to synthesize new methylations. Hereby, the common assumption is, that the enzymes do not show any preference neither for hemi- or unmethylated DNA nor for a DNA-strand. In other words, DNMT3A/B may add a methyl-group to any non-methylated CpG at any DNA-strand. DNMT3L does not catabolize methylation, but increases the binding ability of DNMT3A/B and thus is required for establishing genomic imprinting[1].
The loss of one of the methyltransferases leads to embryonic lethality[1].
For simplification, DNMT3A and DNMT3B will not be further distinguished and called DNMT3.

So far a lot is known about the function of DNMTs, but some properties still remain unexplained:

- Why and where do the DNMTs bind?
- On which conditions does the methylation event depend?
- Which enzymes are able to methylate multiple CpGs in a row without disassociating from the DNA?
- How much de novo and maintenance methylation do DNMT1 and DNMT3A/B perform?

To study the function of DNMTs, a computational, stochastic model is designed and fitted using real biological data.

2.2 Foundations of statistics

Statistics is a mathematical field, dealing with the development of hypotheses, analysis and organization of empirical data. Here, the data are observations of real experiments, often called the sample data. The denotation sample space Ω refers to the collection of possible outcomes of the experiment[16]. Two fields of statistics are distinguishable: descriptive and inferential statistics. Where descriptive statistics summarizes the data without changing it, using statistical methods like mean and standard deviation, inferential statistics is more about analysing data and making predictions based on probability theory[17].

Definition 1 (Probability)**.** *The probability P of an event A is the likelihood of an event to happen. $P(A)$ can take any value between zero and one, where zero describes the impossibility and one the certainty of the event to occur[20].*

Definition 2 (Conditional Probability)**.** *The conditional probability is the probability of an event to happen, given some prior knowledge B:*

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)} \text{ if } P(B) \neq 0 \text{[20].}$$

Definition 3 (Independence)**.** *Two events A and B are independent if $P(A \cap B) = P(A)P(B)$. Which may be written as $P(A \mid B) = \frac{P(A \cap B)}{P(B)} = P(A)$, so if the probability of A does not change compared to the probability of A given B [**Probability**].*

Definition 4 (Random Variable (RV))**.** *A random variable X describes an outcome of an experiment. X is defined as function $X : \Omega \to \mathbb{R}$ [**ProbTheo**].*

Subsequently, we will focus on discrete random variables (RVs), which are a finite set of natural numbers.

Definition 5 (Probability Distribution)**.** *The representation of the probability of each value of a RV is called a probability distribution[**ProbDistri**].*

Definition 6 (Expectation)**.** $E(X) = \sum_{x \in X(\Omega)} x \times P(x)$ *is the expected value of a discrete RV. It may be seen as probability-weighted average of Ω [**Probability**].*

Definition 7 (Variance)**.** *is the squared deviation from the mean.* $V(X) == \sum_{x \in X(\Omega)} (x - E(X))^2 \times P(x)$*[***Probability***]*

Definition 8 (Covariance)**.** $Cov(X, Y) = E((X - E(X))(Y - E(Y)) = E(XY) - E(X)E(Y)$ *is the covariance between two RVs* X *and* Y*. If the value of the covariance is positive, there is a linear relationship between the two variables, otherwise, if the covariance is negative, the variables show opposite behaviour. Finally, the covariance is equal to zero if the variables are uncorrelated[***ProbDistri***].*

Given the definitions one to eight, the data can be described using a stochastic model.
TODO: mean, sd, variance, confidence intervals

2.3 Markov models

A discrete-time Markov model is a stochastic model that fulfils the Markov property. If the next state can only be determined by the current state and not by the previous one, a chain holds the Markov property, also called memoryless property[**Probability**]:

$$P(X_n = x_n \mid X_{n-1} = x_{n-1}, ..., X_0 = x_0) = P(X_n = x_n \mid X_{n-1} = x_{n-1}) \quad (2.1)$$

Markov models can be used to describe a process and its development over time.

A **Markov chain** is the simplest Markov model. Named after Andrey Markov, the first paper about this model was published in 1906[18]. This chain is a sequence of variables $X_1, X_2, ...$ whose outcomes are random, so called random variables (RVs). Each RV has a number of possible outcomes, the state space. Given a set of states $S = \{s_1, ..., s_n\}$ with transitions between the states, p_{ij} is the transition probability from s_i to s_j. And P of size $|S| \cdot |S|$ is the transition matrix containing the transition probabilities between all possible combinations of states.
Let π_0 be the initial distribution of the chain X s.t.

$$\sum_{x \in \pi_0} x = 1,$$

then

$$\pi_1 = \pi_0 \cdot P$$

is the distribution of X at time $t = 1$. We call π the equilibrium distribution; the distribution of X that does not change from one time step to another or formally:
$\pi = \pi \cdot P$.
In general, we distinguish between continuous-time Markov Chains (CTMCs), which act in continuous time and discrete-time Markov Chains (DTMCs) in discrete time.
Markov chains are the most used statistical model for many different real-world processes like for example PageRank[19].

TODO: Ergodicity...?
Hidden Markov Models (HMMs) are an extension of Markov chains and may also be interpreted as simple Bayesian network. There are two kinds of states in this model; hidden states S and observable output states O. Similarly, there is a transition probability matrix between the states in S and a matrix of output probabilities B between the hidden and the output states. Figure 2.3 shows a graphic representation of an HMM.
This model is used, when there is information about the output of a process, but no knowledge about the states of the system.

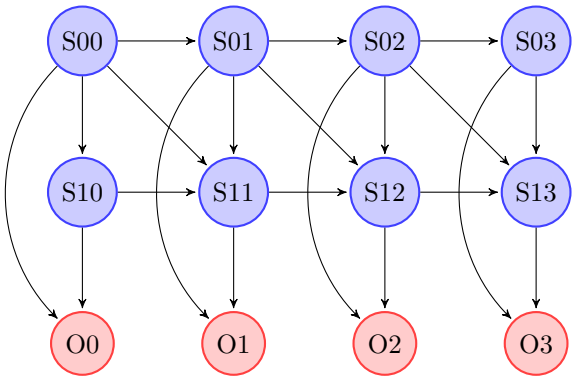

Figure 2.3: *An HMM with eight hidden states(blue) and four emission states(red).*

2.4 Bayesian statistics

By analysing observed data, one is able to develop a stochastic model which represents the process that generated the observed data. Utilizing Bayesian statistical methods, the parameters of the statistical model can be determined. Thereby, Bayesian statistics make use of the Bayes' theorem[20].

Definition 9 (Bayes' Theorem)**.** *Given two events A and B with $P(B) \neq 0$, the conditional probability of the events is given as*
$$P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}.$$

Here, $P(A)$ is the prior probability of A, so our expectation about the process without the knowledge of additional information. $P(A \mid B)$ is called the posterior probability, the probability of A, taking B into account. Finally, $P(B \mid A)$ is the likelihood function; the probability of B, given that A is true.
The likelihood is an important function in Bayesian statistics. By maximizing the likelihood function, the most probable parameter of a model given an observation are decided. The likelihood function, it is defined as

$$\mathcal{L}(\Theta \mid x_0, ..., x_n) = P_\Theta(x_0)...P_\Theta(x_1). \tag{2.2}$$

The process of determining the parameter Θ that maximizes the likelihood function given the data $x_0, ..., x_n$ is called **Maximum Likelihood Estima-**

tion (MLE).

In the following, we describe the two major techniques in Bayesian statistics; Markov Chain Monte Carlo (MCMC) and approximative Bayesian computation (ABC).

The term **MCMC** describes a collection of algorithms which simulate the distribution of a Markov chain after some time t. Once, the initial distribution X_0 and the transition matrix P are known, the simulation works by sampling from these distributions. Therefore, a random number is generated and based on the number, the state of the chain is identified, depending on in which range of the distribution the value is falling.
E.g. assuming, we have three dice, two dice are biased and one regular. We want to compute the probability of throwing six pips. We choose randomly between the biased and the regular dice. The probability distribution of dice selecting is defined by X_0; the probability to dice six pips can be retrieved by considering the transition matrix P.

$$X_0 = \begin{matrix} dice1 & dice2 & dice3 \end{matrix} \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix}$$

$$\mathrm{P} = \begin{matrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ dice1 & \\ dice2 & \\ dice3 & \end{matrix} \begin{bmatrix} 2/15 & 2/15 & 2/15 & 2/15 & 2/15 & 1/3 \\ 1/3 & 1/12 & 1/12 & 1/12 & 1/6 & 1/4 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{bmatrix}$$

Algorithm 1 a Markov Chain Monte Carlo simulation

```
r ← random(0, 1)
for i ∈ len(X_0) do
    if r ≤ X_0[i] then
        dice ← i
        break
    else
        r ← r − X_0[i]
    end if
end for
r ← random(0, 1)
for i ∈ len(P[dice]) do
    if r ≤ P[dice][i] then
        pips ← i
        break
    else
        r ← r − P[dice][i]
    end if
end for
```

The simulation starts by generating the first random number between zero and one. If the value falls into the first interval of X_0, we choose the first dice. Otherwise, if the value is greater than 1/3, but smaller than 2/3, so if the value falls in the second interval, we choose the second dice. If that also does not hold, we select dice three.
After initializing the dice, the rolling of the dice is simulated by generating random variates again. The row, which represents the selected dice, gives the probability distribution of pips. If we roll dice one or two, the probability of throwing six pips is higher than if we rolled dice three. By repeating the simulation multiple times, we get a specific distribution of pips for each combination of selected dice. Comparing a real experiment with multiple throws to our simulated distribution, we are able to predict the dice used in the experiment. Figure 1 shows the pseudo-code of one simulation run of an MCMC.

TODO:forward-backward algorithm?

Algorithm 2 Approximative Bayesian Computation

```
sim ← prior()
if dist(sim, data) < ϵ then
    accept sim
else
    reject sim
end if
```

A second method in computer simulations is **approximative Bayesian computation (ABC)**. In contrast to MCMC, ABC does not need to evaluate the likelihood function, which may be an advantage because the evaluation is computational costly and sometimes not possible. Instead, the similarity of the observed to the real data is measured with a distance function. Hereby, samples are drawn from the prior distribution. If the distance of the sample to the desired value is greater than a threshold ϵ, the sample is rejected, otherwise it is accepted. The posterior distribution results from the set of accepted samples (see figure 4.2 for the pseudo-code)[21].

2.5 Problem

As mentioned in section 2.1, our general goal is to study the function of DNMTs, especially to determine the conditions for methylation activity. We are given the methylation patterns of multiple samples and loci, derived from biological experiments and bisulfite sequencing. For each loci and individual the methylation patterns are known before and after a specific number of replications. These methylation patterns form a distribution, called the methylation pattern distribution. We use Markov models and the given data to model the function of DNMTs and to predict their behaviour. Each state represents the methylation

state of one dyad. Formally,

$$S_i := \text{methylation state of CpG-dyad i} \tag{2.3}$$

$$S_i = \begin{cases} \text{unmethylated} \\ \text{hemimethylated} \\ \text{methylated.} \end{cases} \tag{2.4}$$

Let X be the desired Markov chain and X_0 be the given initial distribution of X. Further, let P be the transition matrix between the states:
$P :=$ methylation probabilities
We want to compute P, such that the distribution of our simulation of the Markov model after n time-steps is as similar as possible to the known distribution X_n.

Problem: Inference of parameters for DNA methyltransferase
Given: initial pattern distribution X_0, pattern distribution at time n X_n
Goal: determine transition matrix P

Recently, different models and computational approaches were studied in order to predict the methylation process. Here we present some of them.

2.6 Related work

One example of such a model to study the methylation activity of DNMT was introduced by **Genereaux et al.**[4] in 2005. Using bisulfite PCR newly synthesized methylations on both strands were discovered, which cannot both result from the malfunction of maintenance methylation. Maintenance methylation occurs only on the daughter strand and independent of which strand is the daughter strand, the origin of the methylation of the other strand is unclear. Thus, the authors assume that de novo methylation exists and develop a model, where μ represents the maintenance methylation rate and δ_p and δ_d the de novo methylation rate at the parent and daughter strand respectively. The methylation state of each CpG-dyad is either M, H or U (methylated, hemimethylated, unmethylated), whereas the state of a single CpG is either m (methylated) or u (unmethylated). Given the methylation rates, the methylation state at time t can be rewritten depending on the previous state like in equations 2.5-2.7.

$$M_t = \mu m_{t-1} + \delta_d \delta_p u_{t-1}, \tag{2.5}$$

$$H_t = \delta_d(1-\delta_p)u_{t-1} + \delta_p(1-\delta_d)u_{t-1} + (1-\mu)m_{t-1}, \tag{2.6}$$

$$U_t = (1-\delta_p)(1-\delta_d)u_{t-1}, \tag{2.7}$$

where $(1-x)$ is the rate of a specific methylation not happening.
By rewriting the equations above, an equations for the equilibrium distribution of the methylation states can be retrieved. Additionally, the likelihood of the states can be computed given the methylation rates.

Fu et al.(2010)[5] used the same model and data from double-stranded DNA to infer the methylation parameters, considering errors in the methylation measuring process. These errors may occur during bisulfite conversion. Regularly, unmethylated cytosines are converted to uracil and methylated cytosines recognized. In the opposite case, unmethylated cytosines are not converted or methylated cytosines falsely converted. The experiments supports the assumption that de novo methylation occurs on both strands. Moreover, the authors come to the conclusion that the methylation rates do not follow their prior distribution, but seems to be location dependant.

Another approach which includes bisulfite conversion errors in the model was introduced by **Arand et al.**[6]. They developed a HMM similar to DTMC proposed by Sontag, Lorincz and Luebeck[22]. The transition matrix of the Markov Chain is composed of three transition matrices; one represents the de novo probability, one the maintenance probability and one the strand segregation probabilities. Based on this model, the equilibrium distribution and the methylation level depending on the methylation probability can be computed. Contrary to Sontag, Lorincz and Luebeck, Arand et al. did not compute the maximal methylation probabilities, but the methylation rates and the influence of bisulfite conversion errors using a maximum likelihood approach.

Recently, **Giehr et al.**[7] developed a similar HMM, which also includes the hydroxylation probability. In the paper the authors distinguish between the occurrence of 5mC and 5hmC and compute their distribution by making use of oxidative bisulfite sequencing. This differentiation was not possible before, using bisulfite-sequencing (BS-seq). Here, a HMM was used with three transition matrices considering the cell division, methylation and hydroxylation probabilities. The observable states of the model are the methylation patterns after the full replication process, whereas the hidden states are the states between single transitions. The parameters estimated using MLE suppose that 5hmC contributes significantly to the demthylation process and that the hydroxylation and methylation level at the loci decreases along the DNA and thus seems to be location-dependant.

The first of three approaches below, covering not only a single CpG at a time for the computation of the methylation parameter, is by **Fu et al. (2012)**[8]. Based on earlier suppositions that the methylation probabilities may rely on neighbouring methylation states, they developed their model to study the processivity of DNMTs. They call an enzyme to work processively if it methylates multiple CpGs in a row without dissociating from the DNA. In total, the model uses four parameter:

$\rho :=$ probability of DNMT to fall off from DNA (dissociation probability)
$\tau :=$ probability of DNMT to bind to DNA (association probability)
$\mu :=$ probability to methylate hemimethylated CpG/CpG-dyad (maintenance methylation probability)
$\delta :=$ probability to methylate unmethylated CpG/CpG-dyad (de novo methylation probability.)

Here, $\rho = 0$ means the enzyme is highly processive, which means it continues its work from one basepair (bp) to another without unbinding. Otherwise, the enzyme can dissociate from the previous bp and bind again at any following bp. The next binding location may be multiple bps apart but may also be the direct neighbour of the position, where the enzyme dissociated. A situation, where the methyltransferase is bound to the DNA at a specific bp and the following bp, may be caused by one of the following events:

1. DNMT works processively (probability $1 - \rho$)
2. DNMT dissociates but reassociates at next bp (probability $\rho * \tau$)

The four parameter are identified for DNMT1, DNMT3A and DNMT3B respectively using the forward-backward algorithm for MCMCs (see section 2.4 for forward-backward algorithm). It is therefore assumed, that DNMT1 works on the daughter strand and DNMT3A/B on the parental as also on the daughter strand.
Additionally included in the model are bisulfite conversion error rates as described in earlier approaches.
In the paper, double-stranded methylation patterns from three loci were used as output states of the HMM. The goal was to assemble the sequence of hidden states, modelled as the association state of the DNMTs. Formally, $P(Q_{ij}, D_{ij} \mid M_{ij}, R^P_{ij}, R^D_{ij})$ is computed, where Q_{ij} and D_{ij} are the methylation states of the parent and daughter strand after replication respectively and M_{ij}, R^P_{ij} and R^D_{ij} are the association states of DNMT1, DNMT3 at the parent and daughter strand. The full transition table holds 9x4 states and is a combination of the four parameters from 2.8.
Two of the three loci used are inactive X (Xi)-linked, namely FMR1 and G6PD. The third locus, LEP, is located on Chromosome seven. In contrast to the autosomal locus, the x-chromosomal loci are highly methylated and consist of large blocks with hemimethylated dyads. The authors suppose that these blocks help to identify the emission probabilities, especially to differentiate between the processive behaviour of the methyltransferases and the dissociation-reassociation event. Consistent to this presumption, in the experiments DNMT1 seems to be highly processive at FMR1 and G6PD, whereas the association parameter are difficult to identify at LEP. Regarding the results of the methylation probabilities, the different loci are conform. DNMT1 performs nearly exclusively maintenance methylation, while no further knowledge may be received from the experiments regarding the methylation probabilities of DNMT3, because the considered confidence intervals of μ and δ are equal to the uniform prior. These results are retrieved by partially restricting the parameter space, which Fu et al. suppose in order to distinguish the activity of DNMT1 and DNMT3.

Based on the findings of Fu et al. (2012), **Bonello et al.**[9] focused on four models and evaluated their advantages and limitations.
First, a location-dependant model was proposed, where the de novo methylation and demethylation rate decrease along the DNA-strand. The de novo methylation probability α and demethylation probability β are defined as:

$$\alpha_i = P(S_i[t] = 1 \mid S_i[t-1] = 0) = exp(-\lambda i), \tag{2.8}$$

$$\beta_i = P(S_i[t] = 0 \mid S_i[t-1] = 1) = exp(\gamma(i - 13)), \tag{2.9}$$

where $S_i[t]$ is the methylation state of the dyad at position i at time t and λ and γ are the parameter of interest.
The other three models are location- and neighbour-dependent and are modelled similar to the model from 2012:

$$P(S_i[t] = 1 \mid S_{i-1}[t] = 0, S_i[t-1] = 0) = 1 - \psi_i(1 - \beta_i). \tag{2.10}$$

Here, ψ_i is the probability that the methylation state at position $i-1$ is different than the one at position i. The three approaches M2, M3, M4 differ only in the way ψ_i is defined.

$$\psi_i = \psi \tag{2.11}$$
$$\psi_i = exp(-\rho i) \tag{2.12}$$
$$\psi_i = exp(\rho(i - 13)) \tag{2.13}$$

In other words, M2 represents a model, where ψ shows no location dependence, and M3 and M4 location dependence following parameter ρ. No model was invented, where neither ψ nor α and β were location dependent as in Fu et al.(2012). The reason for this was, that, looking at the distribution of methylation over the DNA at the O6-methylguanine-DNA methyltransferase (MGMT) promoter, the most methylations seem to be located at the first CpGs and the methylation rate increases towards the end of the promoter region.
The different models were evaluated and the parameter estimated using ABC. It was revealed that model two fitted the four samples from the promoter best. Thus, the methylation state at position i seems to dependent on the preceding methylation state at the whole locus of the same amount, whereas the methylation probability is not stationary over all CpGs[9].

Finally, another idea of neighbour-dependant modelling is to examine not only the preceding CpGs, but to consider CpG-triplets, so to determine methylation probabilities, depending on the methylation state of the preceding and succeeding CpG (**Lück et al.**[10]). Remembering that the DNA is double-stranded, each triplet has six possible methyl-binding positions. Because each of these positions may be methylated or unmethylated, there are $2^6 = 64$ possible methylation patterns for one triplet. Nevertheless, because for each triplet we consider only the change of methylation state of the dyad in the center, there exist only four possible transitions for maintenance and de novo methylation respectively; namely, either none of the neighbouring CpGs, the right neighbour, the left neighbour or both neighbours are methylated. The methylation state of the centered CpG on the parent strand is deciding in case of maintenance methylation. No dependence is between the transition probabilities of the methylation patterns and the upper left and right position of the triplet. The approach also takes bisulfite conversion errors and a strand selection after replication into account.
Four parameters are tracked in total: the maintenance methylation probability μ, the dependency parameters for the left and right neighbour ψ_L, ψ_R and the de novo methylation probability τ. Hereby, the identification of the parameters is performed by MLE. Thereby, the final parameters have the maximal probability to produce the same methylation pattern distribution as our data.
As in [6], three kind of BS-seq data are used. Two of them are knock-out (KO)

data, where either DNMT1 or DNMT3A and DNMT3B are inactivated in the replication process (DNMT1KO, DNMT3A/BKO). The third is wild-type (WT) data, where all enzymes are active.
The results show independence of the succeeding CpG for DNMT1 and DNMT3 and an indicate for a dependence of DNMT3 on the predecessor. Moreover, the dependence of the model on the order of DNMT activity was studied. This means, it was tested if the results changed if the order of DNMTs varies or if the enzymes work alternating at one position compared to the whole strand. Concluding, that the order of DNMT1 and DNMT3A/B matters as the results are different if DNMT1 is not the first enzyme in WT. But the order of DNMT3A and DNMT3B and of the alteration seems not to be significant.

In the following, we describe our model, which follows the approach of [8], introduced earlier. It will be shown that it is possible to distinguish the functions of DNMT1 and DNMT3 and that, using our model, we are able to endorse earlier assumptions about association and methylation rates. Moreover, we compare different computational approaches to estimate the process' parameters.

Chapter 3

Methods

In the following, we present a model for the transmission of methylation patterns by DNA methyltransferases (DNMTs), depending on four parameter, which characterize the function of these DNMTs. Based on that, two computational methods are provided to identify the most realistic way to choose the parameters compared to some real-world methylation data. The aim of this paper is not only to compute the model parameters, but also to compare different approaches to retrieve and evaluate them.

3.1 Model

The model underlies the following procedure:

1. initially a double-stranded, partially methylated DNA sequence with multiple cytosine-phosphate-guanines (CpGs) is given

2. one of the strands serves as template for the methylation of a second yet unmethylated strand

3. the three DNMTs DNMT1, DNMT3a and DNMT3b eventually bind and methylate along the DNA strand

4. the cell divides and step three starts again

Basically, each basepair (bp) is one state of the model, in which the binding state of the DNMTsis kept and if this position is part of a CpG. If the second condition holds, the current methylation state is stored. It is assumed, there are four possible methylation states for one CpG/CpG-dyad. Either none of the cytosines are methylated, then the dyad is unmethylated or the cytosine at the parental or daughter strand is methylated and the respectively other one not, then the CpG is called hemimethylated. In the final case, both cytosines are methylated, which corresponds to a fully methylated dyad. The methylation states are encoded in the following way:

$$m_i := \text{methylation state of dyad i} \tag{3.1}$$

$$m_i = \begin{cases} 0 \text{ unmethylated} \\ 1 \text{ hemimethylated, methylation on parent strand} \\ 2 \text{ hemimethylated, methylation on daughter strand} \\ 3 \text{ fully methylated} \end{cases} \tag{3.2}$$

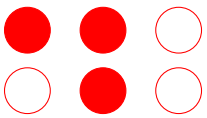

Figure 3.1: *A methylation pattern with three CpGs; each circle represents one CpG; plane red circles are methylated CpGs; the pattern code is 28 and is computed as* $1*4^2+3*4^1+0*4^0 = 16+12+0 = 28$

The different methylation patterns in figure 2.1, whereby subfigure 2.1a corresponds to pattern 0 and so on.
In this way, the full methylation pattern M of a DNA sequence can be encoded as a number in range $4^(L-1)$, where L is the number of CpGs. The code can be interpreted as sum over the methylation states. Hereby the rightmost dyad is the zeroth dyad and all following methylation states are shifted by 2 bits times the enumeration of the dyad (see an example in figure 3.1).

$$M = \sum_{i=0}^{L-1} m_i * 4^{L-1-i}$$

If we are given a set of methylation patterns, the pattern distribution may be represented as dictionary, where the key of each entry is a methylation pattern and its value the frequency of this key in the set of methylation patterns. See Figure **??** for an example.

Further, it is assumed two methyltransferases are active, DNMT1 and DNMT3. For simplification, no other enzymes are included in the model, thus there may be enzymes which influence the methylation process. In earlier experiments it was shown that DNMT1 binds to the daughter strand, whereas DNMT3 is able to bind to both strands (see section 2.1). This knowledge is included by recognizing the binding state of each enzyme to their strands.
Therefore, each state of a CpG can be stored as a combination of four numbers,

X = {0:0.2, 1:0.01, 3:0.02, 11:0.1, 14:0.05, 35:0.01, 36:0.21, 56:0.15, 57:0.05, 63:0.2}

Figure 3.2: *methylation pattern distribution X of a locus with 3 CpGs; each key represents one methylation pattern, the following number the frequency of this pattern; the patterns not listed have frequency zero; assuming, there are 100 samples, the patterns have a absolute frequency of 20, 1, 2, 1, 5, 1, 21, 15, 5 and 2 respectively.*

where the numbers contribute to the methylation state, the binding state of DNMT1 to the new strand and DNMT3 to the old and new strand respectively.

$$\begin{aligned} S_i &:= \text{state of dyad i} \\ S_i &= (m_i, B1_i, B3p_i, B3d_i) \\ m &\in \{0,1,2,3\}^L \\ B1 &\in \{0,1\}^L \\ B3p &\in \{0,1\}^L \\ B3d &\in \{0,1\}^L \end{aligned}$$

Here B1 is the binding state of DNMT1 and B3p the binding state of DNMT3 to the parental strand and B3d the binding state to of DNMT3 to the daughter strand. Zero and one signify an unbound and bound enzyme.

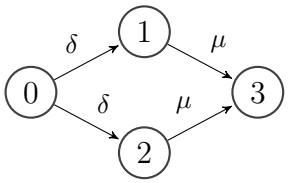

Figure 3.3: *Possible transitions between the methylation patterns of a dyad; the numbers represent the methylation patterns m; each transition shows a possible methylation with the probability denoted on the arrow*

Two kinds of methylations are distinguished in the model. The methylation of a cytosine is called maintenance if the opposite base is yet methylated. Contrary, if the opposite base is unmethylated, the methylation event is called de novo. And the probabilities of each event are denoted by μ and δ, consistent to earlier approaches (2.6). Figure 3.3 shows the resulting possible transitions.

The probabilities of attachment and disattachment of a DNMT are defined as τ and ρ. Therefore, the association length, the number of bps where the enzyme is bound is $1/\rho$ and the probability that the DNMT stays bound at one bp and its neighbour is 1-ρ. This parameter setting is similar to the approach of Fu et al. from 2012.[8] Figure 3.4 visualises the binding states and their probabilities.

Further assumptions for our model are that DNMT1 methylates before DNMT3 as results of [10] suggested this ordering as the most probable and that DNMT1 performs mainly maintenance methylation, whereas DNMT3 is specialized on de novo methylation. Figure 3.5 shows the different combinations of transitions that may lead from one methylation state to another, given the methylation state of the dyad before any methylation activity and the binding states of the three DNMTs. The start methylation state can either be unmethylated or methylated on the parental strand because the newly synthesized strand is initially unmethylated. In this table it is assumed that disassociation and association events have taken place and are excluded in the transition matrix. Because our model does not include demethylation events, a transition from

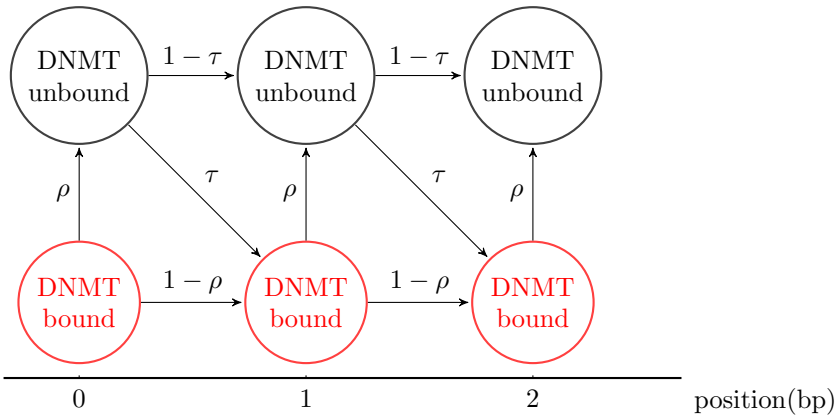

Figure 3.4: *Transitions between the binding states of a DNMT over the DNA; each number represents one bp on the strand; red circles symbolize that DNMT is bound to the DNA at that position, black ones are unbound enzymes; the arrows show possible transitions between the binding states*

methylated to unmehtylated is not possible.

3.2 Simulation

Problem: DNMT simulation
Given: initial methylation pattern distribution X_0, parameters ρ, τ, μ and δ
Goal: generate methylation pattern distribution at time t, X_t

Algorithm 3 Function to draw from initial distribution

```
procedure DRAWINITIALPATTERN(X_0)
    r ← random(0, 1)
    for M, v ∈ X_0 do
        if r > v then
            r ← r − v
        else
            return M
        end if
    end for
end procedure
```

To initialize our model, the methylation pattern M_0 of one DNA-strand at time zero before any replication is needed. We receive one random methylation pattern by drawing from the initial pattern distribution X_0 (see section 3.1 and pseudo-code 3). The number drawn represents the methylation pattern of a double stranded DNA with multiple CpGs. Now, the first cycle of replication starts by cell division. The strands separate and a new, unmethylated strand is synthesized to each already methylated strand. In figure **??** the transitions between the previous double-stranded DNA and the two resulting

	m_i			
$(m_i, B1_i, B3p_i, B3d_i)$	0	1	2	3
(0, 0, 0, 0)	1	0	0	0
(0, 0, 0, 1)	$1-\delta_d$	0	δ_d	0
(0, 0, 1, 0)	$1-\delta_p$	δ_p	0	0
(0, 0, 1, 1)	$(1-\delta_p)(1-\delta_d)$	$\delta_p(1-\delta_d)$	$\delta_d(1-\delta_p)$	$\delta_p\delta_d$
(0, 1, 0, 0)	1	0	0	0
(0, 1, 0, 1)	$1-\delta_d$	0	δ_d	0
(0, 1, 1, 0)	$(1-\delta_p)$	δ_p	0	0
(0, 1, 1, 1)	$(1-\delta_p)(1-\delta_d)$	$\delta_p(1-\delta_d)$	$\delta_p(1-\delta_d)$	$\delta_p\delta_d$
(1, 0, 0, 0)	0	1	0	0
(1, 0, 0, 1)	0	$1-\delta_d$	0	δ_d
(1, 0, 1, 0)	0	1	0	0
(1, 0, 1, 1)	0	$1-\delta_d$	0	δ_d
(1, 1, 0, 0)	0	$1-\mu$	0	μ
(1, 1, 0, 1)	0	$(1-\mu)(1-\delta_d)$	0	$\mu+\delta_d$
(1, 1, 1, 0)	0	$1-\mu$	0	μ
(1, 1, 1, 1)	0	$(1-\mu)(1-\delta_d)$	0	$\mu+\delta_d$

Figure 3.5: *Transition probabilities between the methylation states; m_i on the left side is the methylation state before DNMT activity, either unmethylated(0) or methylated on the parent strand(1), the four rightmost columns represent the four possible methylation states after DNMT activity; the other three parameters in the leftmost column are the binding state of the three different DNMTs; B1 - DNMT1, B3p - DNMT3 at parental strand, B3d - DNMT3 on daughter strand, where 0 means disassociated and 1 associated; later disassociation and association rates are not included in the probabilities; μ is the maintenance methylation probability of DNMT1, δ_p the de novo methylation probability of DNMT3 on the parental strand and δ_d on the daughter strand.*

double-strands are shown. In our simulation we randomly choose one of the two resulting new double-strands with probability 0.5 as shown in pseudo-code 4. Here, M is the previous methylation pattern, $Mnew$ is the chosen new pattern and L is the number of CpGs.

Subsequently, the actual simulation (5) of the methyltransferases starts at the leftmost position of the DNA-chunk and continues till the rightmost position. For each bp it is checked if any DNMT is bound. Initially, all enzymes are not bound and the probability of binding is τ. If any enzyme is bound and if the current position is a CpG, then the DNMT performs de novo methylation with probability δ and maintenance methylation with probability μ. Thereby, de novo methylation is only possible if the opposite strand is unmethylated and maintenance methylation contrary if the opposite strand is methylated. After

m_i	u_i	l_i
0	0	0
1	1	0
2	0	2
3	1	2

Figure 3.6: *methylation states before and after cell division; m_i - before cell division, u_i - the upper strand serves as new parental strand, l_i - lower strand is new parental strand.*

Algorithm 4 Function to simulate cell division

```
procedure cellDivision(M, L)
    r ← random(0, 1)
    Mnew ← 0
    for i ∈ range(L − 1, 0, −1) do
        div ← M div 4^i
        if div = 1 and r ≤ 0.5 then
            Mnew ← Mnew + 4^i
        else
            if div = 2 and r > 0.5 then
                Mnew ← Mnew + 2 × 4^1
            else
                if div = 3 and r ≤ 0.5 then
                    Mnew = Mnew + 4^1
                else
                    if div = 3 then
                        Mnew ← Mnew + 2 × 4^i
                    end if
                end if
            end if
        end if
        M ← M − div × 4^i
    end for
    return Mnew
end procedure
```

a methylation is possibly performed, the transition to the next bp takes place. If an enzyme was bound, the probability of being bound at the next position is $1-\rho$, whereas the event of falling off has probability ρ. Further, the probability of binding of an unbound enzyme is τ again, while the DNMT stays unbound with probability $1-\tau$. We keep track of the binding state of each DNMT until the last bp is reached and store the methylation pattern of the new double strand. The simulation happens first for DNMT1, then DNMT3 and finally for DNMT3 on the parental strand. For the last simulation, the original daughter strand serves as template strand for methylation.

Algorithm 5 Function to simulate DNMT activity

procedure SIMULATEDNMT(M, ρ, τ, μ, δ, len, L) ▷ len := length of DNA chunk in bps
  $bound \leftarrow False$
  $i \leftarrow 0$
  for $pos \in len$ **do**
    if bound or (not bound and random(0,1) $\leq \tau$) **then**
      if isCpG(pos, M) **then** ▷ isCpG decides weather the current bp in M is a CpG
        if (hemi(pos, M) and random(0,1) $\leq \mu$) or (not hemi(pos, M) and random(0,1) $\leq \delta$) **then** ▷ hemi := True if the current CpG in M is methylated on the parental strand, False otherwise
          methylate(pos, M) ▷ methylates the CpG at the current dyad of M
        end if
        $i \leftarrow i + 1$
      end if
      if $random(0, 1) \leq 1 - \rho$ **then**
        $bound \leftarrow True$
      else
        $bound \leftarrow False$
      end if
    end if
  end for
  return M
end procedure

Algorithm 6 Function to generate pattern distribution

procedure SIMULATION($X_0, (\rho 1, \tau 1, \mu 1, \delta 1, \rho 3, \tau 3, \mu 3, \delta 3), len, L, sampleSize, c$) ▷ len := length of DNA chunk in bps, sampleSize := number of simulations, c := number of cell divisions
  $M \leftarrow drawInitialPattern(X_0)$
  for $i \in sampleSize$ **do**
    for $j \in c$ **do**
      $M \leftarrow cellDivision(M, L)$
      $M \leftarrow simulateDNMT(M, (\rho 1, \tau 1, \mu 1, \delta 1), len, L)$
      $M \leftarrow simulateDNMT(M, (\rho 3, \tau 3, \mu 3, \delta 3), len, L)$
      $Minv \leftarrow invert(M)$ ▷ invert converts all 1s to 2s and all 2s to 1s s.t. the original daughter strand serves as parental strand now and vice versa
      $M \leftarrow simulateDNMT(Minv, (\rho 3, \tau 3, \mu 3, \delta 3), len, L)$ ▷ simulates DNMT3 activity on parental strand
      X.append(M)
    end for
  end for
  return X
end procedure

Once one simulation run is finished, the process of drawing from the initial distribution and simulation is repeated multiple times until we receive a methylation pattern distribution.
Because in vivo a cell divides more than one time, we simulate multiple cell divisions by repeating the simulation. Hereby, the methylation pattern after one simulation run is used as base for the next run. Again, one strand is chosen randomly as new template and the other one is yet unmethylated. We use different numbers of cell divisions until a stable pattern distribution is reached, depending on weather we have DNMT1 knock-out (KO), DNMT3a/b KO or wild-type (WT) data. The pseudo-code in 6 shows the whole procedure of methylation pattern distribution simulation.

3.3 Maximum Likelihood Estimation

In order to estimate the program parameter ρ, τ, μ and δ such that they simulate the methylation process best, we need to be able to evaluate the resulting pattern distribution of each parameter combination. Therefore, the pattern distribution X is compared to a given dataset that was retrieved at a specific state of a biological experiment. For evaluation, the following likelihood function $\mathcal{L}$ is used that gives the likelihood that both datasets result from the same parameters θ:

$$\mathcal{L}(\theta) = \sum_{M=1}^{4^L} X_M(\theta)^{N_M}$$

Here, $X_M(\theta)$ is the frequency of pattern with pattern code M in the pattern distribution resulting from the simulation given the parameters θ and N_M is the number of patterns with pattern code i in the measured distribution. Further L is the number of CpGs of this locus. For computational issues, often the log-likelihood function is used instead of the likelihood function. To maximize the likelihood, the simulation is repeated with different θ. The goal function can be written as
$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \mathcal{L}(\theta)$.

3.4 Approximative Bayesian Computation

Another method to determine the best parameters θ for a simulation is approximative Bayesian computation (ABC). In this, each parameter of θ is drawn from a prior distribution. The concluding distribution is compared to a measured using a distance function d. Depending on parameter ϵ, the simulation result is accepted or rejected such that only the distributions with distances smaller than ϵ are stored. Based on the accepted distributions, a new posterior distribution can be build out of which we are able to draw the next θ. After some iterations of the simulation, the parameters of the accepted samples are analysed and assertions about the most likely parameters are possible.
In this approach, we start with a uniform prior to detect the first k samples that are smaller than threshold ϵ. Afterwards, we increase the accuracy of the estimation by generating a distribution based on the already accepted samples and use this posterior for the next iterations. Hereby, each accepted parameter

value builds a distribution, where the distribution mean is the parameter value and the mean deviation from all accepted parameter values of the same parameter the distribution variance. For each parameter t in θ, the distribution π_t out of which is drawn, is chosen with probability $P(\pi_t)$.

$$P(\pi_t) = \frac{\sum_{i \in \Pi_t} d(i) - d(\pi)}{\sum_{j \in \Pi_t} \sum_{i \in \Pi_t} d(i) - d(j)}$$

$\Pi_t :=$ distributions based on parameter t of k accepted samples

$d(i) :=$ distance of distribution i to measured data

Algorithm 7 Generates posterior distributions for four parameters in θ

```
procedure POSTERIOR(thetas, dists)  ▷ thetas := accepted parameters with lowest k distances, dists := distances of best k samples
    for i ∈ {0, 1, 2, 3} do
        Sum ← sum(dists)
        invsum ← 0
        for dist ∈ dists do
            invsum ← invsum + Sum − dist
        end for
        r ← random(0, 1)
        p ← 0
        for j ∈ range(0, len(dists)) do
            p ← p + (Sum − dists[j])/invsum
            if r ≤ p then
                mean ← thetas[j][i]
                for l ∈ range(0, len(dists)) do
                    ABS.append(abs(thetas[j][i]-thetas[l][i]))
                end for
                sd ← mean(ABS)
                break
            end if
            post ← random(mean, sd)
            posterior.append(post)
        end for
    end for
    return posterior
end procedure
```

The pseudo-code in 7 visualizes the new posterior.

We use different distance functions to measure the similarity of the simulated to the given data. The first possibility is to define the distance of to pattern distributions as the absolute value of the difference of the relative frequency of each methylation pattern:

$$d = \sum_{M=0}^{4^L} \mid f_{data}(i) - f_{sim}(i) \mid$$

Hereby, the relative frequency f of a methylation pattern M is the number of occurrences of this pattern over the total number of methylation patterns.
The absolute distance function is compared to another function, which is similar to the Mahalanobis distance function. The Mahalanobis function takes into account that a distribution might have different standard deviations along the different axes because the deviation might be egg-shaped instead of ball-shaped around the center of the distribution. Thus, where the absolute distance and the euclidean distance take the difference of the frequencies, the Mahalanobis distance takes the covariance into account:

$$d = \sqrt{(\vec{f_{data}} - \vec{f_{sim}})^T C^{-1} (\vec{f_{data}} - \vec{f_{sim}})},$$

where C is the covariance matrix and $\vec{f}$ is a vector over all relative frequencies of a distribution.

Algorithm 8 Function that iplements ABC for DNMT simulation

```
procedure ABC(d, ϵ, samples, k, prior, data)                    ▷
d := distanceFunction, ϵ := initial threshold for accepted distances, samples
:= number of generated samples, k := number of accepted samples, prior :=
prior distribution, data := measured distribution to compare with
    for i ∈ samples do
        if len(thetas) < k then
            theta ← prior()
        else
            theta ← posterior(thetas, dists)
        end if
        sim ← simulation(theta)
        dist ← d(sim, data)
        if dist < ϵ then
            thetas.append(theta)
            dists.append(dist)
            distributions.append(sim)
            if len(thetas) > k then
                Max ← max(dists)
                thetas.del(Max)
                dists.del(Max)
                distributions.del(Max)
                ϵ ← mean(dists)
            end if
        end if
    end for
    return (thetas, dists, distributions)
end procedure
```

Once, k samples were accepted, ϵ is refined by setting its new value to the mean distance of k accepted samples and the sample with the worst distance is discarded. After one more simulation value was accepted, we recompute ϵ again and reject another worse sample such that there are always k accepted samples (see pseudo-code 8). Given k distributions with the lowest distance to

the measured distribution, the parameter confidence intervals can be computed by the mean parameter values and their standard deviations.

Chapter 4

Results

We use two distinct approaches to compute the parameter that specify the function of DNA methyltransferases (DNMTs), especially their binding and methylation probabilities. In our computer simulation we focus on DNMT1 and DNMT3. The enzyme activity is modelled using a Hidden Markov Model (HMM), that stores the binding state of all methyltransferases and the methylation state for each (cytosine-phosphate-guanine (CpG))-position. Three scenarios are considered: DNMT1 knock-out (KO), where only DNMT3 is active; DNMT3a/b KO, where DNMT3 is absent and wild-type (WT), where both enzymes are involved in the methylation process. For WT it is assumed the order of enzyme activity is DNMT1 on daughter strand, DNMT3 on daughter strand and DNMT3 on parental strand. Supposing the DNA is not unmethylated before replication, an initial methylation pattern is given by WT data. Thereupon, the iterations of the simulation generate a methylation pattern distribution that is compared to a measured distribution in order to evaluate it.
The measured datasets are derived using hairpin-bisulfite sequencing. Three loci are considered; two repetitive elements: major Satellites (mSat) and IAPLTR1, LTR-retrotransposons (IAP) and an alpha feto protein gene (Afp). For more details see [6].

4.1 Maximum Likelihood Estimation

locus	restrictions	ρ	τ	μ	δ	likelihood
mSat		0.894	0.276	0.782	1	4757
mSat	$\rho = 1$	1	0.275	0.565	1	4704
mSat	$\tau = 1$	≈ 1	1	0.17	0.279	4713
Afp		1	0	0.809	0.857	1692
IAP		0	0	0.028	0.566	2675

Figure 4.1: *Parameter for minimal negative log-likelihoods of three loci for DNMT1KO.*

locus	restrictions	ρ	τ	μ	δ	likelihood
mSat		0.234	1	0.736	0.426	3496
mSat	$0.8 \leq \mu \leq 1$	≈ 0	≈ 1	0.8	0.443	3430
mSat	$\rho = 1$	1	0.929	0.885	0.417	3429
Afp		≈ 0	0.488	0.366	≈ 0	2659
IAP		≈ 0	0.994	0.837	0.496	2494

Figure 4.2: *Parameter for minimal negative log-likelihoods of three loci for DNMT3a/bKO.*

In the first method, the program parameters were estimated using the likelihood function from chapter **??**. First, the locus mSat will be considered here as it is the one with the fewest CpGs. Let us focus on the KO data first, where either DNMT1 or DNMT3 are active. Looking at the first row of table **??** and **??**, where the results of Maximum Likelihood Estimation (MLE) are displayed for locus mSat and DNMT3 and DNMT1 respectively, one can see the differences in the function of the two enzymes as their parameters are clearly distinguishable. Where the association probability is very high and the dissociation probability rather low for DNMT1, the values are vice versa for DNMT3. This means that DNMT1 will always bind to the DNA and the few times where it falls off, it is very likely that it will bind again. Where the association length, the number of basepairs (bps) where an enzyme is bound, is very high for DNMT1, the same length is small for DNMT3. DNMT3 is not very likely to bind to the DNA and will fall off after a short wile. Moreover, the de novo methylation probability is higher for DNMT3, which is consistent with our suggestions. Whereas the maintenance methylation probability is between 0.7 and 0.8 for both enzymes, DNMT1 will perform more maintenance methylation than DNMT3 because its association length is higher. If we multiply the maintenance methylation probability of DNMT3, μ_3, with its association probability, we get the real value of μ_3 because the enzyme needs to be bound to methylate. $\mu_3 \times \tau_3$ is equal to 0.216 and thus clearly smaller than the maintenance methylation probability of DNMT1.

We evaluate the likelihood function further by investigating the borders of the function. We thus compare the value of the likelihood after each parameter was set to zero and one to the present minimum. Further, during the MLE computation, it was striking that sometimes some parameters were difficult to identify because the range where this parameter maximizes the likelihood was wide as visible in figure **??**. Therefore, another approach was to fix an interval for the parameters, which may maximize the likelihood. The best results for the parameter restrictions for mSat are shown in row two and three in table **??** and **??**. For both methyltransferases, the likelihood is maximized by setting ρ to one. τ and δ do not change significantly compared to the result without parameter restriction for DNMT3, while μ decreases by 0.217. For DNMT1 τ decreases and μ increases a bit such that all parameters besides δ are very high. Regularly, the interpretation of a high ρ would be that the enzyme falls off at each bp, but in with case, if τ is next to one too, the DNMT will immediately

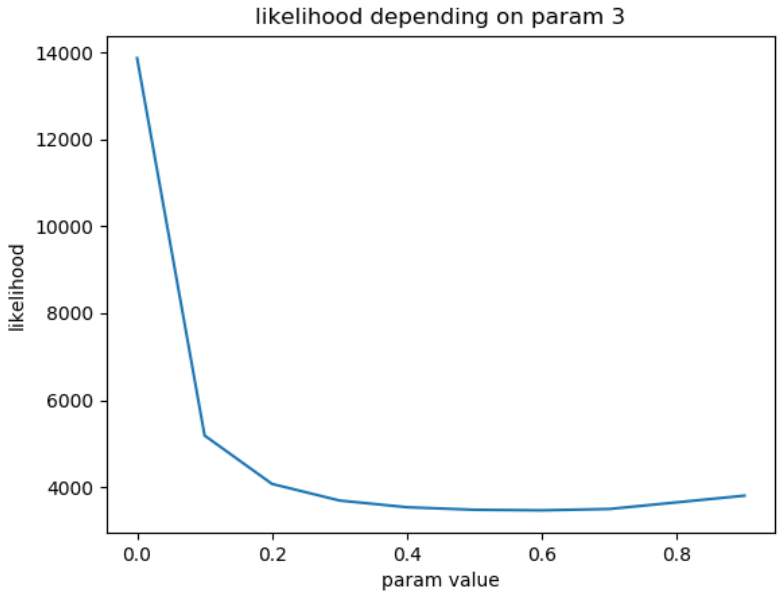

Figure 4.3: *Negative log-likelihood depending on parameter δ; the other three parameter values are fixed.*

bind again. In this model, the enzyme can be interpreted as very processive. We are not able to distinguish between a low disassociation probability and a very high dis- and reassociation probability as both combinations lead to the event that the methyltransferase is bound to one bp and its successor. Thus, DNMT1 is interpreted as very processive and the result with $\rho = 1$ and $\tau \approx 1$ is similar to the previous result. Another very likely scenario for DNMT3 is, if τ was set to one, ρ is approximately one, but the methylation probabilities decrease compared to the approach without parameter restriction. The interpreting is different here, because the methylation activity is low. DNMT3 seems to work processive in this case but methylation events occur rather rarely. In the third scenario of DNMT1 and locus mSat, μ was restricted to the interval [0.8,1]. Compared to row one of table **??**, both parameter estimations are very similar again; only the disassociation probability is even lower. Both results suggest a processive behaviour of DNMT1 with high methylation probabilities.

4.2 Approximative Bayesian Computation

Chapter 5

Conclusions

todo

Appendix A

Regulation

Bibliography

[1] B. Jin et al. "DNA Methylation". In: *Genes Cancer* 2.6 (2011). DOI: 10.1177/1947601910393957.

[2] K. Janitz and M. Janitz. "Handbook of Epigenetics". In: Academic Press, 2011. Chap. 12. DOI: 10.1016/B978-0-12-375709-8.00012-5.

[3] W. Reik and J. Walter. "Genomic imprinting: parental influence on the genome". In: *Nature Reviews Genetics* 2 (2001).

[4] D.P. Genereaux et al. "A population-epigenetic model to infer site-specific methylation rates from double-stranded DNA methylation patterns". In: *PNAS* 102.16 (2005).

[5] A.Q. Fu et al. "STATISTICAL INFERENCE OF TRANSMISSION FIDELITY OF DNA METHYLATION PATTERNS OVER SOMATIC CELL DIVISIONS IN MAMMALS". In: *Ann Appl Stat* 4.2 (2010).

[6] J. Arand et al. "In Vivo Control of CpG and Non-CpG DNA Methylation by DNA Methyltransferases". In: *plos genetics* (2012). DOI: 10.1371/journal.pgen.1002750.

[7] P. Giehr et al. "The Influence of Hydroxylation on Maintaining CpG Methylation Patterns: A Hidden Markov Model Approach". In: *plos computational biology* (2016). DOI: 10.1371/journal.pcbi.1004905.

[8] A.Q. Fu et al. "Statistical Inference of In Vivo Properties of Human DNA Methyltransferases from Double-Stranded Methylation Patterns". In: *plos one* (2012).

[9] N. Bonello et al. "Bayesian inference supports a location and neighbour-dependent model of DNA methylation propagation at the MGMT gene promoter in lung tumours". In: *Journal of theoretical biology* 336 (2013). DOI: 10.1016/j.jtbi.2013.07.019.

[10] A. Lueck et al. "A Stochastic Model for the Formation of Spatial Methylation Patterns". In: *Computational Methods in Systems Biology. CMSB 2017. Lecture Notes in Computer Science*. Ed. by Koeppl H. Feret J. Vol. 10545. Springer, Cham, 2017. ISBN: 978-3-319-67470-4. DOI: 10.1007/978-3-319-67471-1_10.

[11] B. Weinhold. "Epigenetics: The Science of Change". In: *Environment Health Perspective* 114.3 (2006). DOI: 10.1289/ehp.114-a160.

[12] Y. Li and T.O. Tollefsbol. "DNA methylation detection: Bisulfite genomic sequencing analysis". In: *Methods Mol Biol.* 791 (2011). DOI: 10.1007%2F978-1-61779-316-5_2.

[13] Y. Huang et al. "The Behaviour of 5-Hydroxymethylcytosine in Bisulfite Sequencing". In: *plos one* (2010). DOI: 10.1371/journal.pone.0008888.

[14] D.-Q. Shi et al. "New Insights into 5hmC DNA Modification: Generation, Distribution and Function". In: *Front Genet.* 8.100 (2017). DOI: 10.3389%2Ffgene.2017.00100.

[15] M. G. Goll et al. "Methylation of tRNAAsp by the DNA Methyltransferase Homolog Dnmt2". In: *Science* 311.5759 (2006). DOI: 10.1126/science.1120976.

[16] J.-W. Romeijn. "Philosophy of Statistics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2017. Metaphysics Research Lab, Stanford University, 2017.

[17] H. Keone. *Introduction to Statistics*. CreateSpace Independent Publishing Platform, 2014. ISBN: 978-1502424624.

[18] C.M. Grinstead and J.L. Snell. *Introduction to Probability*. American Mathematical Soc., 1997. ISBN: 0821807498.

[19] P.A. Gagniuc. "Markov Chains: From Theory to Implementation and Experimentation". In: NJ: John Wiley and Sons, 2017. ISBN: 978-1-119-38755-8.

[20] J.M. Bernardo. *Bayesian Statistics*.

[21] B.M. Turner and T. Van Zandt. "A tutorial on approximate Bayesian computation". In: *Journal of Mathematical Psychology* 56 (2012).

[22] L.B. Sontag et al. "Dynamics, stability and inheritance of somatic DNA methylation imprints". In: *journal of theoretical biology* 242.4 (2006). DOI: 10.1016/j.jtbi.2006.05.012.