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-methylations. Hereby, the gene expression is inactivated by binding of methyl-groups to 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). The specific function of DNMTs is not fully determined.

In the following, the methylation of DNA is modelled using an Markov Chain Monte Carlo (MCMC) algorithm and parameters are estimated by Maximum Likelihood Estimation (MLE). Alternatively, parameter estimation is performed with an implementation of the Approximative Bayesian Computation (ABC) method. Moreover, a method to compare distributions of methylation patterns is provided.
We find differences in the function of the different DNMTs that are consistent with current biological findings. Thus, 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

ABC Approximative Bayesian Computation ii
CGI CpG island 3
CpG cytosine-phosphate-guanine ii
CTMC Continuous-time Markov Chain 5
DNMT DNA Methyltransferase ii
DTMC Discrete-time Markov Chain 5
HMM Hidden Markov Model 5
MCMC Markov Chain Monte Carlo ii
MLE Maximum Likelihood Estimation ii
PCR polymerase chain reaction 9
RV Random Variable 5

Chapter 1

Introduction

Chapter 2

Background

2.1 DNA methylation

Epigenetics is the name of the science that 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.[1]

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 CpGs, where a cytosine nucleotide (C) is followed by a guanine nucleotide (G) in the DNA sequence.[2] Whereas the majority of CpGs in mammals are methylated, so called CpG island (CGI) are rather unmethylated. The CGIs are associated to gene regulation as they are often located at the promoter region of genes. Hereby, methylation inhibits the gene expression; ensuring that changes in the methylation pattern of the DNA are effecting diseases like cancer.[3] Hypomethylation of repeat elements for example results in an unstable DNA and may increase the risk of cancer; as also the hypermethylation of cancer suppressor genes.[2] 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. [4]

The transfer of the methyl group to the DNA is performed by DNMTs. Five different DNMTs are distinguished in mammals: DNMT1, DNMT2, DNMT3A, DNMT3B and DNMT3L.[2]
DNMT1 is known as maintanance methyltransferase as its activity is associated with 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. DNMT1 has a preference for hemimythelated DNA, which means that one of the opposite CpGs is methylated and the other unmethylated. Subsequently, DNMT1 transfers the methylation by methylating the positions on

the daughter strand that are methylated at the same position on the parental strand.[2]
DNMT2 is negligible if human DNA methylation is considered, because it methylates small RNA at the anticodon loop.[5]
DNMT3A and DNMT3B are de novo methyltransferases, which work during the early embryonic development to synthesize new methylations. Hereby, 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.[2]
The loss of one of the methyltransferases leads to embryonic lethality.[2]

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 deassociating 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.

Todo: bisulfite PCR

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.[6] Two sections 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.[7]

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

Definition 2 (Conditional Probability)**.** *The conditional probability is the probability of an event to happen, given some prior knowledge K:*
$P(E \mid K) := P(E \cap K)/P(K)$ *if* $P(K) \neq 0$.

Definition 3 (Independence)**.** *Two events A and B are independent if* $P(A \mid B) = P(A)$*, so if the probability of A does not change compared to the probability of A given B.*

If two events are independent it holds $P(A \cap B) = P(A) + P(B)$ otherwise $P(A \cap B) = P(A) + P(B) - P(A \cup B)$

Definition 4 (Random Variable (RV))**.** *A random variable X describes an outcome of an experiment.*

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

Given the Definitions one to five, the data can be described using a stochastic model.

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

A **Markov chain** is the simplest Markov model. This chain is a sequence of variables $X_1, X_2, ...$ whose outcomes are random, so called 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 X_0 be the initial distribution of the chain X s.t. $\sum_{x \in X_0} x = 1$, then $X_1 = X_0 * P$ 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: $\pi * P = \pi$.
In general, we distinguish between Continuous-time Markov Chains (CTMCs), which act in continuous time and Discrete-time Markov Chains (DTMCs) in discrete time.
Hidden Markov Models (HMMs) are an extension of Markov chains. 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 2.1 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.

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.

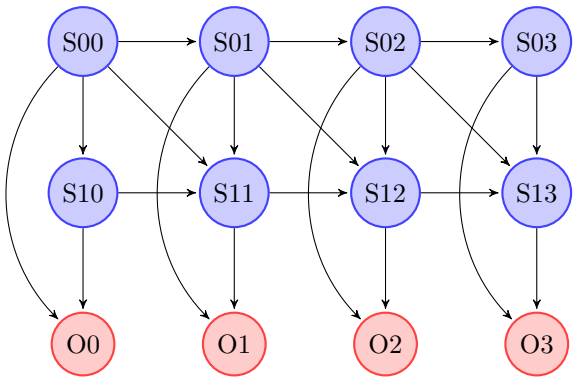

Figure 2.1: *An HMM with eight hidden states(blue) and four emission states(red).*

Thereby, Bayesian statistics make use of the Bayes' theorem.[8]

Definition 6 (Bayes' Theorem)**.** *Given two events A and B with $P(B) \neq 0$, the conditional probability of the events is given as $P(A \mid B) = (P(B \mid A) * P(A))/P(B)$.*

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 observed. Because of the importance of the likelihood function, it is also defined as

$$L(\Theta \mid x_0, ..., x_n) = P_\Theta(x_0) * ... * P_\Theta(x_1). \tag{2.1}$$

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

In the following, we describe the two major techniques in Bayesian statistics; MCMC and 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.

$r \leftarrow random(0, 1)$
for $i \in len(X_0)$ **do**
  if $r \leq X_0[i]$ **then**
    $dice \leftarrow i$
    break
  else
    $r \leftarrow r - X_0[i]$
  end if
end for
$r \leftarrow random(0, 1)$
for $i \in len(P[dice])$ **do**
  if $r \leq P[dice][i]$ **then**
    $pips \leftarrow i$
    break
  else
    $r \leftarrow r - P[dice][i]$
  end if
end for

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

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

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 **??** shows the pseudo-code of one simulation run of an MCMC.

A second method in computer simulations is **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, the

sample is rejected, otherwise it is accepted. The posterior distribution results from the set of accepted samples (see figure **??** for the pseudo-code).[**ABC**]

$sim \leftarrow prior()$
if $dist(sim, data) < \epsilon$ **then**
    accept sim
else
    reject sim
end if

2.5 Problem

In the context of DNMTs, Markov models are used to simulate the function of these enzymes, making use of results from biological experiments. Here, few is known about the properties of methyltransferases. Under which condition does the enzyme bind to the DNA and when does it methylate? But the methylation patterns before and after the catabolism of the DNMTs are known. From the methylation pattern distribution, a sequence of methylation states can be retrieved. Each methylation (unmethylated, hemimethylated, methylated) of one CpG-dyad is an output state, whereas the binding state and methylation conditions of the DNMT at each CpG are hidden states. Thus the traversing probabilities between the hidden states P are equal to the probabilities of the DNMTs to bind/fall of and the output probabilities represent the different methylation probabilities.
Given the output sequence

$$O_i := \text{methylation state of CpG-dyad i} \tag{2.2}$$

$$O_i = \begin{cases} \text{unmethylated} \\ \text{hemimethylated} \\ \text{methylated,} \end{cases} \tag{2.3}$$

we are able to determine the most likely sequence of S
$S_i :=$ state of DNMT at CpG i
Moreover, our goal is to determine

$$P := \text{probability of DNMT to change binding state} \tag{2.4}$$

$$B := \text{methylation probabilities.} \tag{2.5}$$

Figure 2.2 shows the four kind of methylation patterns.

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

Figure 2.2: *Kind of methylation patterns; 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*

2.6 Related work

One example of such a model to study the methylation activity of DNMT was introduced by Genereaux et al. in 2005. Using bisulfite polymerase chain reaction (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 on which strand is the daughter strand, the origin of the mathylation 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 nove 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 2.8.[**Genereaux**]

$$M_t = \mu * m_{t-1} + \delta_d * \delta_p * u_{t-1} \tag{2.6}$$

$$H_t = \delta_d * (1 - \delta_p) * u_{t-1} + \delta_p * (1 - \delta_d) * u_{t-1} + (1 - \mu) * m_{t-1} \tag{2.7}$$

$$U_t = (1 - \delta_p) * (1 - \delta_d) * u_{t-1}, \tag{2.8}$$

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.[**Genereaux**]

Fu et al. 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 its model was introduced by Arand et al..[**Wolf**] They developed a HMM similar to DTMC proposed by Sontag, Lorincz and Luebeck.[**Sontag**] The transition matrix of the Markov Chain is composed of three transition matrix'; 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 Lubeck, 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.[**Wolf**]

Recently, Giehr et al. developed a similar HMM, which also includes the hydroxylation probability. An idea, which we will not enlarge upon in the following.[**Giehr**]

Chapter 3

Methods

Chapter 4

Evaluation

Chapter 5

Discussion

todo

Appendix A

Regulation

Bibliography

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

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

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

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

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

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

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

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